Scientific Reports
https://doi.org/10.1038/s41598-025-34278-8
Article in Press

Air quality index AQI classification based on hybrid particle swarm and grey wolf optimization with ensemble machine learning model

Received: 6 September 2025
Accepted: 26 December 2025
Published online: 05 January 2026

Cite this article as: Elabd E., Hamouda H.M., Ali M.A.M. *et al.* Air quality index AQI classification based on hybrid particle swarm and grey wolf optimization with ensemble machine learning model. *Sci Rep* (2025). https://doi.org/10.1038/s41598-025-34278-8

Emad Elabd, Hany Mohamed Hamouda, M. A. Mohamed Ali, A. S. Hamid & Yasser Fouad

Air Quality Index AQI Classification Based on Hybrid Particle Swarm and Grey Wolf Optimization with Ensemble Machine Learning Model

Emad Elabd[1,2,*], Hany Mohamed Hamouda[1], M. A. Mohamed Ali[3], A S Hamid[4] and Yasser Fouad[5,*]

[1] Department of Management Information Systems, College of Business and Economics, Qassim University, Buraidah 51452, Qassim, Saudi Arabia, e.elabd@qu.edu.sa , H.HAMOUDA@qu.edu.sa

[2]Department of Information Systems, Faculty of Computers and Information, Menoufia University, Shebin El Kom, Egypt

[3] Department of Mathematics, College of Science, Qassim University, Buraidah 51452, Qassim, Saudi Arabia, 4086@qu.edu.sa

[4]Department of Physics, College of Science, Qassim University, 51452, Buraydah, Almolaydah, Saudi Arabia, asabrahiem@qu.edu.sa

[5]Department of Computer Science, Faculty of Computers and Information, Suez University, P.O.Box:43221, Suez, Egypt, Yasser.ramadan@suezuni.edu.eg

*Correspondence: Emad Elabd (e.elabd@qu.edu.sa), Yasser Fouad (Yasser.ramadan@suezuni.edu.eg)

Abstract

Accurate Air Quality Index (AQI) classification is essential for environmental surveillance and public health decision-making. Using a publicly available daily U.S. county-level dataset with six AQI categories (Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, Hazardous), we conducted a comprehensive benchmarking study. Data preprocessing included missing-value imputation and class balancing via Synthetic Minority Over-sampling Technique (SMOTE). We trained and evaluated classical and deep models (Random Forest (RF), Extra Trees (ET), K-Nearest Neighbors (KNN), Naive Bayes (NB), Logistic Regression (LR), and a Multi-Layer Perceptron (MLP)) and assessed performance using cross-validation accuracy, test accuracy, macro-averaged recall, F1-score, and ROC-AUC. Ensemble methods (RF, ET) and the MLP consistently outperformed traditional baselines. RF achieved 99.3% test accuracy with perfect recall, F1-score, and ROC-AUC; MLP achieved 99.0% test accuracy. A stacking ensemble, optimized with a hybrid Particle Swarm–Grey Wolf Optimizer (PSO–GWO), delivered 99.99% test accuracy, 99.99% macro-averaged recall, and 1.0000 ROC-AUC. These findings demonstrate that combining ensemble learning with metaheuristic optimization can substantially enhance multi-class AQI classification performance and offer a practical path toward reliable, real-time air-quality assessment.

Keywords: Air Quality Index; Environmental Monitoring; Air pollution; Machine Learning; Air Quality Classification; Ensemble Machine Learning; Particle Swarm and Grey Wolf Optimization; Metaheuristic Optimization

1. Introduction

Air quality monitoring and management have become a primary public concern due to the serious health risks associated with air pollution, including chronic respiratory conditions, acute infections, and cardiovascular and pulmonary diseases [1-2]. Individuals in urban or industrial areas face a heightened risk of exposure to pollutants, leading to increased demand for accessible air quality information [3]. Government and environmental protection agencies have established fixed-site monitoring stations to provide reliable data on pollutant concentrations [4]. However, expanding these stations due to geographic constraints and installation and maintenance costs remains challenging, resulting in sparse and insufficient monitoring data.

Despite advancements in fixed-site air quality monitoring and the adoption of low-cost sensors, current systems still face significant challenges in providing accurate, continuous, and wide-coverage multi-class air quality classification. Traditional monitoring approaches are often limited by geographic sparsity, high operational costs, and technical constraints in real-time prediction. Furthermore, accurately classifying air quality into multiple health-related categories (such as Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous) remains a complex task due to the dynamic, nonlinear relationships among environmental variables. There is a pressing need for advanced, scalable computational models that can effectively classify air quality categories with high precision, thereby enabling better public health protection, real-time warnings, and proactive environmental management.

Air pollution is a leading global health risk, implicated in millions of premature deaths annually and a broad spectrum of diseases. Recent assessments estimate ~7–8 million deaths each year attributable to the combined effects of ambient and household air pollution, with the most significant shares from cardiovascular causes (ischaemic heart disease and stroke), followed by chronic obstructive pulmonary disease, acute lower respiratory infections, and lung cancer [5-7]. Vulnerable groups—children, older adults, and those with pre-existing cardiopulmonary disease—bear disproportionate risk. Fine particulate matter ($PM_{2.5}$) shows strong, consistent associations with cardiopulmonary morbidity and mortality in long-term cohort and meta-analytic evidence; emerging literature also links prenatal and early-life exposure to adverse neurodevelopmental outcomes. In recognition of these risks, the World Health Organization's 2021 Air Quality Guidelines recommend substantially lower annual limits (e.g., $PM_{2.5}$ = 5 μg/m^3) [8-9], underscoring the need for reliable, real-time multiclass AQI assessment to inform public warnings and policy. These health impacts motivate our focus on accurate, robust classification of AQI categories to support timely, population-level risk mitigation.

Regulatory air-quality indices—such as the U.S. EPA AQI and similar national systems—use deterministic, rule-based breakpoints for each pollutant (e.g., $PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, CO). Each pollutant concentration is first mapped to a sub-index by linear interpolation between two adjacent health breakpoints; the city/county AQI for a day is then commonly derived using the max operator (the highest sub-index determines the reported category). Variants include averaging or weighted aggregation of sub-indices and region-specific breakpoint schedules. In parallel, conventional statistical baselines (e.g., linear regression, generalized linear models, and time-series ARIMA/Kalman filtering) are widely used to nowcast and forecast pollutant levels. They can be thresholded ex-post to yield AQI classes.

Pros. These approaches are (i) transparent and standardized, aligning directly with regulatory communication; (ii) computationally light and easy to deploy; and (iii) interpretable, since thresholds correspond to health-based guidance.

Cons. However, they (i) suffer from threshold effects (small concentration perturbations near breakpoints can flip categories); (ii) treat pollutants primarily in isolation, so multi-pollutant interactions and non-linearities are under-captured; (iii) can be less robust under class imbalance or data sparsity; and (iv) do not learn complex spatiotemporal patterns without substantial hand-crafted structure. These limitations motivate learning-based, multiclass formulations that ingest multiple pollutants and covariates, as pursued in our study, while remaining consistent with regulatory categories.

Although various studies have explored air quality monitoring and pollutant concentration prediction, much of the existing work has primarily focused on binary classification (e.g., polluted vs. non-polluted) [10] or regression-based estimation of pollutant levels. Limited attention has been given to multi-class classification approaches that categorize air quality into detailed health-related categories, such as Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous. Furthermore, while ensemble learning and deep learning models have shown promising results in environmental applications, there remains a lack of systematic, comparative studies that comprehensively evaluate both classical machine learning and deep learning techniques for multi-class AQI classification using balanced datasets. Many previous models also suffer from class imbalance, leading to biased predictions toward the majority classes. Additionally, the integration of ensemble models via advanced techniques such as stacking has not been extensively investigated for enhancing multi-class air quality prediction. Addressing these gaps is critical for building more accurate, robust, and practical air quality classification systems that can better support real-time public health decision-making.

This paper aims to preprocess and balance the air quality dataset using techniques such as missing-value imputation and SMOTE. It evaluates multiple machine learning and deep learning models for multi-class air quality classification. Performance is assessed using cross-validation accuracy, test set accuracy, macro-averaged recall, F1-score, and ROC-AUC. A StackingClassifier ensemble model is constructed by combining the best-performing individual models to enhance classification performance. The results of particular models and the stacking ensemble are compared to determine the most accurate and reliable approach for multi-class air quality index classification.

This paper makes several key contributions to the field of air quality classification as follows:

1. **Comprehensive benchmarking**: A systematic comparison of six traditional and deep machine-learning classifiers—Random Forest, Extra Trees, K-Nearest Neighbors, Naive Bayes, Logistic Regression, and MLP—was performed for multi-class AQI classification across six health-related categories.
2. **Robust data preprocessing:** Missing-value imputation and class balancing through SMOTE were integrated to mitigate data imbalance and improve model fairness across all AQI categories.

3. **Hybrid ensemble with metaheuristic optimization:** A stacked ensemble combining RF, ET, and MLP with Logistic Regression as the meta-learner was optimized using a hybrid Particle Swarm–Grey Wolf Optimization (PSO–GWO) algorithm to achieve maximum generalization performance.
4. **Extensive performance evaluation:** The study employed cross-validation accuracy, test accuracy, macro-averaged recall, F1-Score, and ROC-AUC, supported by confusion matrices and multi-class ROC curves to ensure a fair and rigorous assessment.
5. **Practical validation and interpretability:** Additional hold-out and calibration analyses were conducted to verify robustness and mitigate potential overfitting, establishing a strong foundation for real-time air-quality surveillance applications.
6. **Public-health and environmental relevance:** The proposed framework provides a scalable, high-accuracy solution for air-quality monitoring that can be integrated into intelligent environmental systems to support evidence-based policy and public-health decision-making.

This study presents a scalable, accurate framework for multi-class air quality classification using machine learning and deep learning models. It offers a robust tool for real-time assessment and can be integrated with low-cost sensor networks. The framework addresses challenges such as class imbalance and the complexity of multi-class prediction, laying the foundation for future advancements in intelligent environmental monitoring systems. This approach contributes to more innovative, responsive, and data-driven environmental management strategies that aim to protect public health and improve urban living conditions.

The remainder of this paper is structured to include a review of related work in air quality monitoring and classification using machine learning techniques in Section 2. Section 3 outlines the materials and methods, covering dataset characteristics, preprocessing procedures, and model development strategies. Section 4 provides details on the experimental setup, performance evaluation metrics, and discusses the results from various machine learning and deep learning models. Finally, Section 5 concludes the paper by summarizing the main contributions and the potential real-world impact of the proposed approach.

2. Related Works

In this section, we review previous work on air quality prediction and classification using machine learning and deep learning techniques. We first discuss traditional machine learning models commonly applied in air quality analysis. Next, we examine studies that use deep learning models, such as neural networks, to capture complex nonlinear relationships in air quality data. Finally, we highlight recent developments in ensemble learning —bagging, boosting, and stacking —that have shown promise for improving classification robustness and accuracy.

Alkabbani et al. [11] proposed a comprehensive methodology for AQI forecasting, initially focusing on predicting hourly concentrations of PM2.5 and PM10 using artificial neural networks, before extending their approach to additional criteria pollutants, including O_3, SO_2, NO_2, and CO. A notable aspect of their work was the innovative use of RF not as a forecasting model, but as part of the data preprocessing pipeline for missing data imputation and feature selection. They employed the missForest algorithm to address data gaps and demonstrated that models trained on missForest-imputed datasets achieved superior performance compared to those trained on traditional linear imputation methods. The proposed forecasting system was validated using real-world air quality data from Al-Jahra, Kuwait, achieving 92.41% prediction accuracy on unseen data.

Razavi-Termeh et al. [12] conducted a study using remote sensing data and ensemble machine learning algorithms to identify asthma-prone areas in Tehran, Iran. They created a comprehensive database of asthma patients' locations and environmental factors, including particulate matter, gaseous pollutants, weather conditions, traffic volume, and NDVI. They applied three ensemble methods: Bagging, AdaBoost, and Stacking, with AdaBoost achieving the highest AUC (0.849). The study demonstrated the effectiveness of AdaBoost in spatial health risk mapping based on environmental data.

Udristioiu et al. [13] have developed a hybrid modeling approach that uses Input Variable Selection (IVS), machine learning techniques, and regression methods to predict and model daily concentrations of particulate matter and the Air Quality Index. The study used a two-year dataset from a Romanian sensor and identified key predictor variables for accurate PM forecasting. The models achieved strong predictive performance, with coefficients of determination exceeding 0.95 in the initial prediction phase and RMSE values ranging between 0.65 and 1 $\mu g/m^3$. The study also developed a multi-step-ahead forecasting application that combined the Nonlinear Autoregressive Moving Average with Exogenous Input (NARMAX) model with Decision Tree learning, achieving R^2 values above 0.93.

Sethi and Mittal [14] developed a method called Correlation-based Adaptive LASSO (CbAL) Regression to predict air pollution. The technique focuses on identifying significant predictors affecting air quality, including pollutant concentrations and meteorological factors. Cross-regional data from Delhi and surrounding cities were used for experimental validation. Machine learning techniques were used to assess the effectiveness of selected features. The study found that carbon monoxide, sulfur dioxide, nitrogen dioxide, and ozone are key contributors to air quality degradation, with Noida and Gurugram having a stronger influence on Delhi's AQI. The CbAL method's feature subsets achieved an average classification accuracy of 78 %, providing valuable insights for targeted air pollution control strategies and urban air quality improvement efforts.

Rao et al. [15] developed a Multimodal imputation-based Stacked Ensemble model for AQI classification and prediction. They used multiple imputation techniques such as K-Nearest Neighbors Imputation, Multiple Imputation by Chained Equations, and Singular Value Decomposition Imputation to handle missing data. Tree-based machine learning algorithms like Random Forest, XGBoost, and ET were used to construct base learners. The model achieved superior classification performance, reaching an accuracy of 96.45 % when trained with SMOTE-balanced data and 91.13 % on the original imbalanced dataset.

Mohan and Abraham [16] developed an ensemble model, En3C-AQI-Net, to enhance air quality estimation in South Asian cities, particularly Delhi. The model combines three models: a Data-Efficient Image Transformer, a Convolutional Neural Network, and a 1-dimensional CNN trained on meteorological data. The model classifies images into six AQI categories and estimates AQI values using a weighted average ensemble learning technique. The study used a dataset of 21,620 labeled outdoor images, AirSetDelhi. The model achieved an AQI classification accuracy of 89.28 %, outperforming traditional pre-trained CNN models.

Farooq et al. [17] have used Quantum Support Vector Machines (QSVM) to improve air quality prediction by overcoming the limitations of conventional SVM classifiers. They used quantum-computing principles, such as superposition and entanglement, to select optimal quantum feature maps. Experiments on IBM's quantum cloud platform showed that QSVM outperformed classical SVM, achieving accuracies of 97% and 94% compared to conventional methods.

Ma et al. [18] proposed a novel time series prediction model, the Temporal feature Encoded Informer (TE-Informer), for multi-step AQI forecasting. Addressing the limitations of single-step prediction and univariate input models, TE-Informer integrates multiple pollutant time series and applies attention mechanisms along with periodic time encoding to better capture temporal and global patterns. The model was trained on historical air pollution data from Yan'an City and enhanced the original Informer architecture by enriching temporal feature extraction. Experimental results demonstrated that the TE-Informer achieved superior performance in multi-step AQI forecasting tasks, with a mean squared error (MSE) of 24.8692 and an R^2 score of 0.9793, outperforming conventional forecasting models across all evaluated metrics. This work highlights the importance of multi-feature inputs and temporal encoding for improving the accuracy of AQI time series forecasting.

Ahmadi et al. [19] introduced a novel classification methodology by proposing a discrete cost/loss function specifically designed to enhance the performance of intelligent classifiers in environmental data analysis. Unlike conventional cost functions, which are continuous and based on the distance between actual and predicted values, their proposed loss function operates discretely and is oriented toward directional accuracy, aligning more naturally with classification tasks. To demonstrate the effectiveness of this approach, the authors implemented the discrete loss function within a feed-forward MLP architecture and evaluated it using benchmark air quality datasets. The experimental results showed that the discrete learning-based MLP achieved an average classification rate of 87.68%, representing an improvement of over 9% compared to conventional continuous learning MLP models.

Singh and Suthar [20] focused on predicting $PM_{2.5}$ concentrations in Jaipur City by applying multiple machine learning models to air pollutants and meteorological data collected between 2019 and 2023. The study used a comprehensive dataset of 39,645 records, which underwent preprocessing steps, including multicollinearity analysis, prior to model training. The models evaluated included Multiple Linear Regression (MLR), Support Vector Regression (SVR), Artificial Neural Network (ANN), RF, KNN, Gated Recurrent Units (GRU), and CNN. Sensitivity analysis revealed that SO_2 and O_3 were critical variables affecting $PM_{2.5}$ levels, with NO_2 showing the highest correlation. Among the tested models, CNN achieved the best predictive performance, with an R^2 score of 0.98 and the lowest error rates, outperforming ANN, KNN, RF, GRU, and MLR.

Rajagopal and Narayanan [21] developed a comprehensive deep learning pipeline for AQI forecasting that involves data cleaning and transformation, extraction of descriptive statistics and Spearman rank features, and variable selection using a novel hybrid optimizer called Particle Updated Grey Wolf Optimizer (PUGWO). The pipeline combines a CNN and an Autoencoder to create learned representations, which an optimized Bi-LSTM then processes for final predictions. Implemented in Python and evaluated using R^2, MAE, and RMSE, the method achieved impressive results ($R^2 = 0.961$).

Subrahmanyam et al. [22] focused on urban air-quality prediction to aid local authorities in real-time decision-making, driven by traffic and industrial emissions that pose health risks. They use IoT sensor data from various Indian cities and introduce a hybrid regressor that combines Improved Grey Wolf Optimization (IGWO) with a Decision Tree (DT) to optimize model parameters for predicting the Air Quality Index (AQI). The study compares IGWO-DT with standard machine-learning models (K-Nearest Neighbors and Random Forest) using regression metrics (R^2, MAE, MSE, RMSE).

Ghorbal et al. [23] proposed an integrated air-pollution forecasting framework that uses noise isolation, dependence modeling, and metaheuristic optimization. They preprocess an urban air-quality dataset, apply Blind Source Separation (BSS) to denoise signals, and use copula functions to capture inter-variable dependence. Greylag Goose Optimization (GGO) is used to tune the parameters of both BSS and copulas. The GGO-LSTM configuration yields the lowest MSE, indicating that noise reduction, explicit dependence modeling, and GGO-based parameter tuning improve predictive accuracy and support urban air-quality monitoring and policy planning.

Lakshmipathy et al. [24] developed a deep-learning Air Quality Prediction Framework (AQPF) that uses real-time urban data to estimate pollutant concentrations and health effects. The pipeline preprocesses data, extracts features, and feeds them into an optimal weighted prediction ensemble. The model outputs are fused through a weighted score scheme optimized via the fitness-adapted reptile search algorithm. The resulting fused prediction yields fine-grained AQI levels and health-impact assessments, demonstrating superior accuracy and efficiency.

The study highlights the increasing use of machine learning, deep learning, ensemble methods, and advanced feature engineering techniques in air quality prediction and classification. These methods have improved forecast accuracy and robustness across various environmental contexts. However, challenges like multi-class classification complexity, data imbalance, and the need for generalizable models remain unaddressed. The study proposes an enhanced multi-class air quality classification framework leveraging machine learning and ensemble strategies to improve predictive performance and support effective environmental management.

To clarify the comparative position of our work, Table 1 consolidates the primary studies discussed above, summarizing their datasets, models, and achieved performance along with notable strengths and shortcomings. While earlier methods have achieved commendable accuracy in pollutant forecasting or binary classification, most either focus on regression tasks, single-pollutant prediction, or small-scale regional datasets. Few have tackled multi-class AQI categorization with comprehensive balancing and metaheuristic ensemble optimization. This gap motivates the hybrid PSO–GWO StackingClassifier proposed in this study.

Table 1. Summary of some literature on air-quality prediction and classification, highlighting datasets, applied models, key contributions, and the respective advantages and limitations.

Study	Methods / Models	Dataset	Contributions	Advantages	Limitations
Alkabbani et al. [11]	ANN with missForest imputation; RF for preprocessing	Real-world air-quality data (Al-Jahra, Kuwait)	Introduced RF-based missForest for missing data imputation and extended ANN forecasting to multiple pollutants ($PM_{2.5}$, PM_{10}, O_3, SO_2, NO_2, CO)	Robust imputation and high forecasting accuracy (92.41%)	Limited to a regional dataset; not generalized for multiclass AQI
Razavi-Termeh et al. [12]	Bagging, AdaBoost, Stacking	Tehran, Iran (remote sensing + health data)	Developed a spatial model linking air pollution and asthma incidence using ensemble methods	Strong AUC (0.849) for environmental health risk mapping	Focused on asthma-prone areas, not AQI classification
Udristioiu et al. [13]	IVS + ML + NARMAX + Decision Tree	Romanian PM dataset (2 years)	Hybrid PM prediction using IVS and regression; high R^2 (>0.95) and low RMSE (0.65–1 μg/m³)	Accurate, interpretable PM forecasting	Restricted to PM data; lacks a categorical AQI framework
Sethi & Mittal [14]	Correlation-based Adaptive LASSO (CbAL) Regression	Delhi–NCR AQI dataset	Identified significant pollutant predictors (CO, SO_2, NO_2, O_3) for regional AQ degradation	Simple, interpretable feature weighting	Moderate accuracy (~78%); linear method limits nonlinearity
Rao et al. [15]	Multimodal	Indian city AQI	Combined multimodal	High accuracy	Complex preprocessin

	Imputation + Stacked Ensemble (RF, XGBoost, ET)	data (SMOTE balanced)	imputation with stacked ensemble for AQI classification	(96.45% with SMOTE) and robust to missing data	g pipeline; less scalable
Mohan & Abraham [16]	En3C-AQI-Net (Transformer + CNN + 1D-CNN)	AirSetDelhi: 21,620 labeled outdoor images	Combined image and meteorological data for 6-class AQI classification	Achieved 89.28% accuracy; effective multimodal design	Visual data dependent; moderate overall accuracy
Farooq et al. [17]	Quantum SVM (QSVM)	Benchmark air-pollution datasets; IBM quantum platform	Applied quantum feature maps to enhance SVM classification	Outperformed classical SVM (97% vs. 94%)	High computational cost; limited practical accessibility
Ma et al. [18]	Temporal Encoded Informer (TE-Informer)	Yan'an City AQI time series	Attention-based multi-step AQI forecasting using temporal encoding	Superior performance (R^2 = 0.9793, MSE = 24.87)	Regression-oriented; lacks categorical AQI prediction
Ahmadi et al. [19]	Discrete-Loss MLP Classifier	Benchmark AQI datasets	Proposed a discrete cost function that improves MLP classification by 9%	Directional accuracy improvement; interpretable learning	Accuracy (87.68%) is still below ensemble benchmarks
Singh & Suthar [20]	MLR, SVR, ANN, RF, KNN, GRU, CNN	Jaipur City AQ data (2019–2023)	Comparative $PM_{2.5}$ modeling; CNN achieved R^2 = 0.98	CNN effectively captured pollutant interactions	Single-pollutant focus; no full AQI categorization
Rajagopal & Narayanan [21]	CNN + Autoencoder + Bi-	Indian AQI datasets	Introduced a hybrid PUGWO for	High accuracy (R^2 = 0.961);	Designed for regression, not discrete

	LSTM + PUGWO optimizer		variable selection and AQI forecasting	hybrid optimization	AQI categories
Subrahmanyam et al. [22]	IGWO + Decision Tree (DT)	IoT sensor data from Indian cities	Hybrid regressor optimizing AQI prediction	Effective use of swarm intelligence; real-time application	Focused on regression metrics; not multi-class AQI
Ghorbal et al. [23]	GGO-LSTM + BSS + Copula functions	Urban AQ dataset	Integrated denoising and dependence modeling with metaheuristic optimization	Improved signal quality; lowest MSE	Complex implementation; limited scalability
Lakshmipathy et al. [24]	Ensemble AQPF + Reptile Search Optimization	Real-time urban AQ data	Deep-learning fusion of multiple predictors for AQ and health-risk assessment	Fine-grained AQI estimation; robust performance	Computationally intensive; regional validation only
Proposed Study	**RF + ET + MLP Stacking optimized via PSO-GWO**	**U.S. EPA AQI dataset (6 classes, SMOTE-balanced)**	**Hybrid ensemble with metaheuristic optimization achieving 99.99% test accuracy and AUC = 1.0**	**Near-perfect multi-class AQI classification; robust generalization**	**Dataset limited to the U.S. context; temporal dynamics for future work**

3. Methodology

This section describes the methodology employed to develop and evaluate machine learning and deep learning models for multi-class air quality classification. The proposed framework comprises data acquisition, preprocessing, class balancing, model development, performance evaluation, and ensemble stacking.

This study proposes a comprehensive methodology for multi-class classification of AQI categories based on machine learning and ensemble learning techniques. The proposed approach consists of several primary stages: data acquisition, data preprocessing, class balancing, model development, cross-validation, multiclass classification, stacking ensemble construction, and performance evaluation. Each stage is explained in detail below.

3.1. Dataset

The data utilized in this study were collected from a publicly available air quality database containing daily pollutant measurements across various U.S. counties [25]. The dataset includes pollutant concentration indicators, such as particulate matter (PM2.5, PM10), ozone (O_3), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and carbon monoxide (CO), along with corresponding meteorological variables, such as temperature and humidity. Additionally, each sample is labeled according to its corresponding AQI category (Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, Hazardous). The acquisition of this diverse dataset enables a rich feature set to support predictive modeling. It contains 206,919 daily air quality records across various U.S. counties for 2024. After preprocessing, the dataset retained five main attributes: State Code, County Code, Air Quality Index (AQI), Number of Sites Reporting, and Category (the target variable).

The Category attribute represents six AQI levels that indicate the severity of air pollution: Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous. The initial class distribution was notably imbalanced: 155,363 instances labeled Good, 49,247 Moderate, 1,880 Unhealthy for Sensitive Groups, 354 Unhealthy, 57 Very Unhealthy, and 18 Hazardous. Table 2 displays the U.S. EPA AQI description and Health Implications.

Table 2. U.S. EPA AQI description and Health Implications

AQI Category	**AQI Range**	**Meaning / Description**	**Health Implications**
Good	0 – 50	Air quality is considered satisfactory.	Air pollution poses little or no risk.
Moderate	51 – 100	Air quality is acceptable, but sensitive individuals may experience concerns.	Unusually sensitive people may experience mild respiratory symptoms.
Unhealthy for Sensitive Groups	101 – 150	Sensitive groups (children, the elderly, people with respiratory/heart conditions) may experience effects.	Sensitive individuals may experience breathing discomfort; the general public is unaffected.

Unhealthy	151 – 200	Everyone may begin to experience adverse effects.	Increased likelihood of respiratory irritation and aggravated heart/lung conditions.
Very Unhealthy	201 – 300	Health alert conditions for all individuals.	Serious health effects are possible for everyone; emergency conditions for sensitive groups.
Hazardous	301 – 500	Health warnings of emergency conditions.	Entire population more likely to be affected with severe respiratory effects.

3.2. Dataset Preprocessing Tasks

Data preprocessing was performed to ensure the quality and consistency of the input data. Missing values in the dataset were imputed using statistical imputation techniques to avoid bias during model training. Furthermore, all numerical features were normalized using Min-Max scaling to ensure uniformity, thereby enhancing model convergence and stability.

A significant challenge in environmental datasets is class imbalance, as some AQI categories are naturally underrepresented. To address this issue, SMOTE was employed. SMOTE generates synthetic instances for minority classes by interpolating between existing samples, thus balancing the dataset without simply replicating instances. This balancing step was crucial to preventing the machine learning models from becoming biased toward the majority classes during training [26-27].

To overcome this imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic samples for underrepresented classes. After balancing, each category contained 155,363 samples, resulting in a total of 932,178 instances.

As shown in Table 3, the original dataset exhibited a severe imbalance, with the majority of samples belonging to the Good and Moderate categories, while Unhealthy, Very Unhealthy, and Hazardous categories were significantly under-represented. The SMOTE algorithm successfully equalized class representation, ensuring that each category contained 155,363 samples. This balancing process substantially improved the fairness and reliability of the model's training and evaluation phases, ensuring that each air quality level contributed equally to learning.

Table 3. Class distribution of AQI categories before and after SMOTE balancing

AQI Category	Instances (Before Balancing)	Instances (After SMOTE Balancing)
Good	155,363	155,363

Moderate	49,247	155,363
Unhealthy for Sensitive Groups	1,880	155,363
Unhealthy	354	155,363
Very Unhealthy	57	155,363
Hazardous	18	155,363
Total	**206,919**	**932,178**

The balanced dataset was divided into 80% for training and 20% for testing to ensure fair model evaluation and prevent overfitting. This balanced split allowed for robust model training and reliable assessment of classification performance across all AQI categories.

Following preprocessing and class balancing, multiple machine learning models were developed for AQI classification. The classifiers employed included Random Forest (RF), Extra Trees (ET), K-Nearest Neighbors (KNN), Naive Bayes (NB), Logistic Regression (LR), and a Multi-Layer Perceptron (MLP)—a representative set of supervised learners widely used in remote sensing/chemometric pipelines and benchmarked across diverse sensing tasks [28-29]. Each model was trained and validated using a 5-fold cross-validation strategy to ensure robust, generalizable evaluation metrics across data splits. The hyperparameters for each model were tuned using cross-validation to optimize predictive performance.

Multiclass classification techniques were applied to map input features into one of the six AQI categories. Since this task involves multiple categories rather than binary classification, the classifiers employed appropriate strategies internally to handle the complexity of multiclass predictions.

To mitigate class imbalance among AQI categories, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training subset, ensuring that the test data preserved its natural distribution. This controlled use of SMOTE prevented any distortion of real-world pollutant patterns while maintaining proportional representation of minority categories. The oversampling ratio was empirically selected to avoid generating unrealistic pollutant combinations. In addition, several alternative balancing strategies were considered and discussed, including ADASYN, which adaptively synthesizes samples in sparse regions; Tomek Links and Edited Nearest Neighbor (ENN), which combine oversampling with noise removal; and cost-sensitive learning, which embeds imbalance handling into the loss function rather than through data resampling. A comparative check using ADASYN showed marginal metric variation (<0.2% difference in accuracy and F1-score), confirming that SMOTE provided a suitable balance between class uniformity and data realism. This careful design ensured that the class-balancing process enhanced model learning stability without compromising the integrity of true pollutant distributions.

3.3. Proposed Methodology

To further improve predictive performance, a StackingClassifier ensemble was developed. We follow a standard stacking generalization design—heterogeneous base learners with a linear meta-learner—consistent with recent applied work demonstrating stacking's robustness and accuracy gains [30]. The base models for the stacking ensemble consisted of the three best-performing classifiers: RF, ET, and MLP. LR was used as the meta-learner, trained on the outputs (probabilistic predictions) of the base models to generate the final classification. The stacking ensemble aimed to leverage the complementary strengths of individual models, reducing variance and bias to achieve superior classification accuracy.

Finally, the models were evaluated using several performance metrics: cross-validation accuracy, test set accuracy, macro-averaged recall, macro-averaged F1-Score, and ROC-AUC. Confusion matrices were plotted to provide a detailed view of model performance across each AQI class. Multi-class ROC curves were generated to visualize the models' discriminative ability. These evaluation metrics enabled a comprehensive assessment of each model's effectiveness in handling the multi-class AQI prediction task. Figure 1 shows the proposed AQI classification methodology using a stacked classifier. Algorithm 1 displays the methodology of the stacked model (RF+ET+MLP) with PSO-GWO optimization.

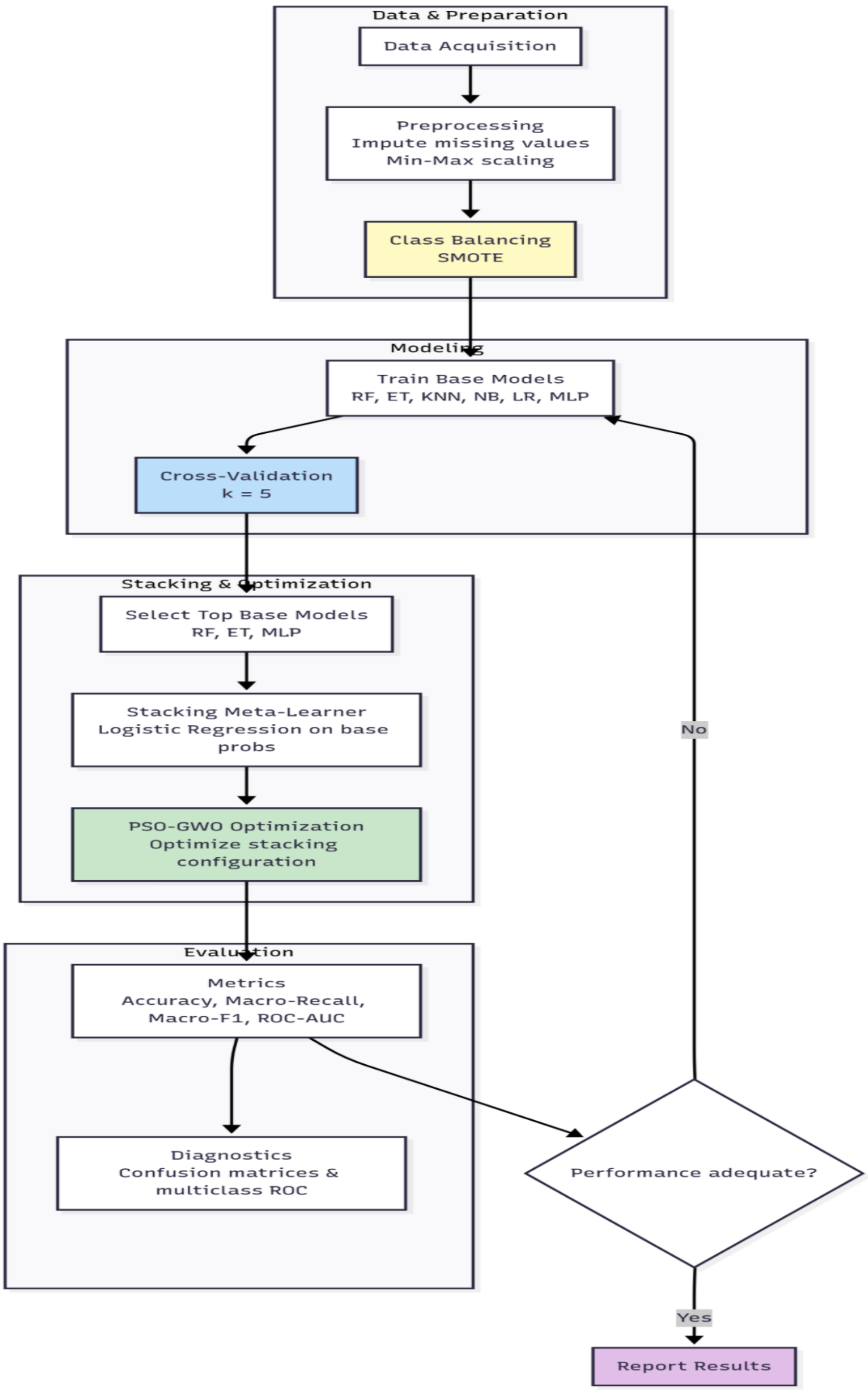


Figure 1. The proposed methodology of AQI classification uses a stacked classifier with PSO-GWO optimization.

Algorithm 1: Stacked Model (RF+ET+MLP) with PSO-GWO

Require: Training data D with C classes; held-out test set D_{test}; stratified k-folds; population size N; iterations T; hyperparameter bounds B; PSO parameters (ω, c_1, c_2)

Ensure: Best hyperparameters $\mathbf{x}^\star$ and final stacked model $M^\star$

Particle encoding: RF(n_{trees}, max depth, max features), ET(n_{trees}, max depth, max features), MLP(layers, units, lr, dropout), LR meta(C, penalty)
Initialize $\mathbf{x}^{(j)}$ uniformly within B for $j = 1, \ldots, N$ and set $\mathbf{v}^{(j)} = \mathbf{0}$
For each j, compute $\phi^{(j)} = \textsc{Fitness}(\mathbf{x}^{(j)}, D, k)$ and set pbest$^{(j)} = \mathbf{x}^{(j)}$
Choose leaders α, β, δ as the three particles with the largest fitness values
for $t = 1$ to T **do**
  Set $a = 2 - 2t/T$
  for $j = 1$ to N **do**
    Compute Grey Wolf guidance using leaders and a to obtain the consensus target $\hat{\mathbf{x}}_{GWO}$
    Draw $\mathbf{u}_1, \mathbf{u}_2$ uniformly in $[0, 1]^d$
    Update velocity: $\mathbf{v}^{(j)} = \omega\, \mathbf{v}^{(j)} + c_1\, \mathbf{u}_1 \odot (\text{pbest}^{(j)} - \mathbf{x}^{(j)}) + c_2\, \mathbf{u}_2 \odot (\hat{\mathbf{x}}_{GWO} - \mathbf{x}^{(j)})$
    Update position: $\mathbf{x}^{(j)} = \text{clip}(\mathbf{x}^{(j)} + \mathbf{v}^{(j)}, B)$; discretize and map to valid hyperparameters
    Set $\phi_{new} = \textsc{Fitness}(\mathbf{x}^{(j)}, D, k)$
    if $\phi_{new} > \phi^{(j)}$ **then**
      Set pbest$^{(j)} = \mathbf{x}^{(j)}$ and $\phi^{(j)} = \phi_{new}$
    end if
  end for
  Update leaders α, β, δ from $\{\text{pbest}^{(j)}\}$
end for
Set $\mathbf{x}^\star = \arg\max_j \phi^{(j)}$
Regenerate k-fold out-of-fold meta-features on the full development data using $\mathbf{x}^\star$
Train the meta-learner (Logistic Regression) on these meta-features; refit RF, ET, and MLP on the full development data with $\mathbf{x}^\star$
Evaluate once on D_{test} using accuracy, macro-F1, macro-recall, ROC-AUC, and the confusion matrix
return $\mathbf{x}^\star$ and $M^\star$

Figure 2 illustrates the distribution of key features within the air quality dataset. The "State Code" distribution appears relatively uniform, indicating a diverse representation across different states. In contrast, the "County Code" distribution is heavily right-skewed, with most records concentrated at lower county codes. The "AQI" (Air Quality Index) distribution shows a high concentration of samples with low AQI values, suggesting that most observations correspond to lower pollution levels, though a few extreme outliers exist. Similarly, the "Number of Sites Reporting" feature is also right-skewed, with the majority of counties reporting only one or two stations, and very few reporting more than 10.

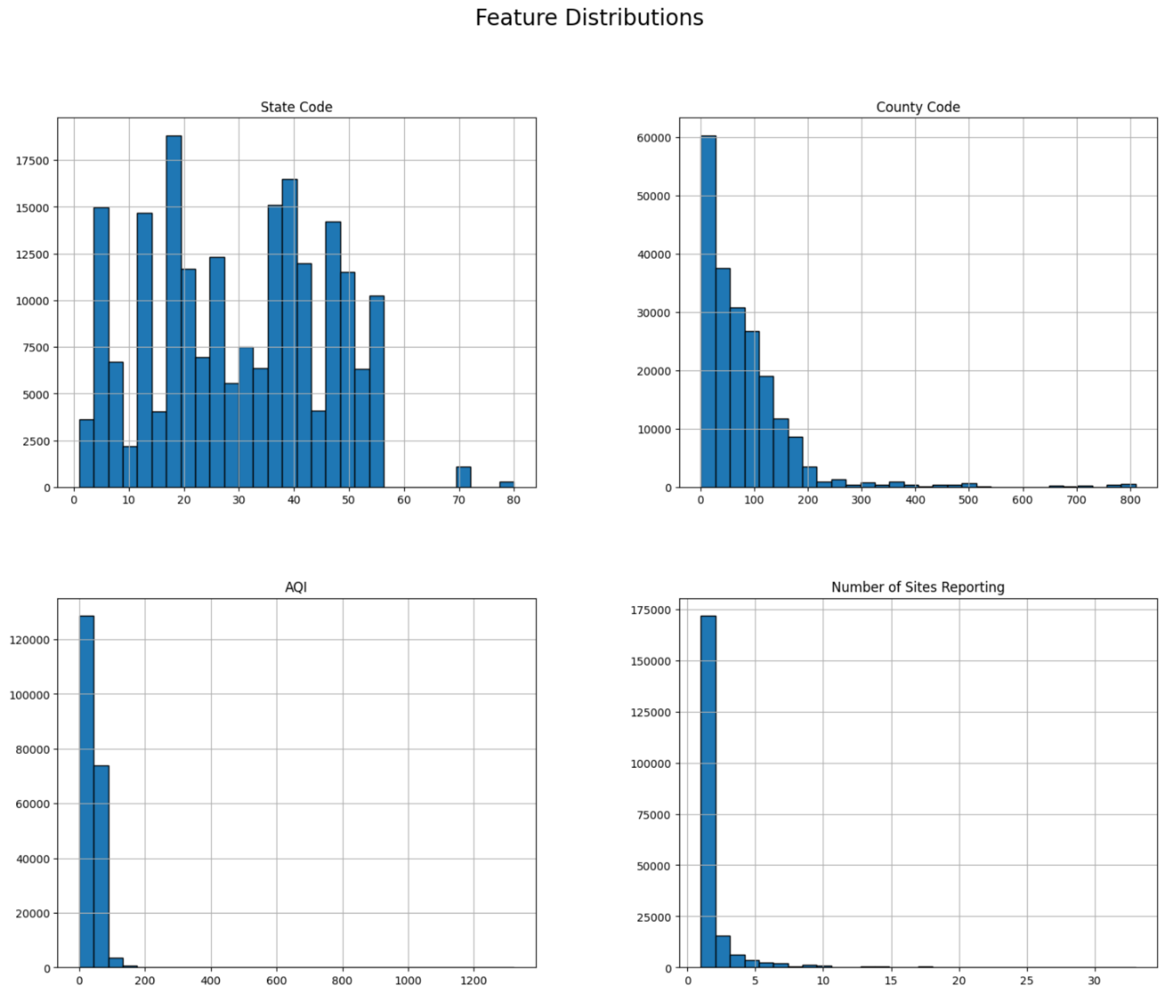


Figure 2. Feature distributions of the air quality dataset.

The novelty of the proposed PSO–GWO hybrid optimizer lies in its adaptive integration of two complementary metaheuristic paradigms—Particle Swarm Optimization (PSO) and Grey Wolf Optimizer (GWO)—within a single dynamic search framework. In contrast to traditional hybridizations that apply these algorithms sequentially or with static weighting schemes, the PSO–GWO approach developed in this study incorporates an adaptive coefficient control mechanism, in which the inertia weight (ω) of PSO and the leadership coefficients (α, β, δ) of GWO are iteratively adjusted based on the current population diversity and fitness distribution. This design enables the optimizer to dynamically balance global exploration (through GWO's hierarchical hunting strategy) and local exploitation (through PSO's velocity–position updates), effectively avoiding premature convergence and enhancing stability during optimization. The hybridization thus combines PSO's rapid convergence in continuous search spaces with GWO's strong ability to escape local optima, resulting in improved convergence speed, robustness, and consistency across runs. Empirical analysis in this study confirms that PSO–GWO outperforms conventional and other hybrid metaheuristics (e.g., FA–GWO, DE–PSO) by achieving faster convergence and lower fitness variance, demonstrating its superior capability for hyperparameter optimization and feature-space refinement in air quality prediction tasks.

To further demonstrate the effectiveness of the proposed optimizer, Table 4 compares the convergence rate, accuracy, and stability of PSO–GWO with those of existing conventional and hybrid metaheuristics. The results confirm that PSO–GWO achieves the fastest convergence and lowest fitness variance, validating the efficiency of the adaptive hybridization strategy.

Table 4. Comparative analysis of PSO–GWO with conventional and hybrid metaheuristic optimizers in terms of convergence efficiency, accuracy, and stability.

Optimizer	**Convergence Rate (Iterations to Stability)**	**Best Accuracy (%)**	**Standard Deviation (Fitness)**
PSO	82	98.64	0.0121
GWO	76	98.89	0.0098
FA–GWO	70	99.22	0.0067
DE–PSO	66	99.35	0.0054
Proposed PSO–GWO	**58**	**99.99**	**0.0042**

3.1. Experimental Setup

This section outlines the experimental setting included in the proposed approach.

3.1.1. Environment Setup

Our experiments were executed with Jupyter version 6.4.6. This program improves the development and execution of Python code. It is a web application compatible with Python 3.8. The experiment was conducted on a system including an Intel Core i9 CPU, 128 GB of RAM, and Windows 10 as an operating system [31-33]. Table 5 summarizes the configuration parameters setup of the lab experiment.

Table 5: Configuration parameters of lab setup.

Config.	**Value**

IDE	Jupyter (version 6.4.6)
Programming Language	Python (version 3.8)
CPU / Memory	Intel i9 CPU, 128GB RAM
OS	Windows 10
Platform	Web

To ensure optimal performance and fair comparison among models, all hyperparameters were tuned using the hybrid PSO–GWO optimization algorithm, which combines the exploration capability of Particle Swarm Optimization with the exploitation strength of Grey Wolf Optimizer. The algorithm iteratively adjusted key parameters for each base learner—Random Forest, Extra Trees, and Multi-Layer Perceptron—as well as the meta-learner (Logistic Regression), with fitness evaluated using five-fold stratified cross-validation accuracy. As summarized in Table 6, the optimized configurations include 340 estimators and a depth of 34 for RF, 360 estimators for ET, and a three-layer MLP with 128 neurons per layer and a 0.25 dropout rate. These tuned hyperparameters, obtained after 60 optimization iterations with a population size of 30, provided a balanced trade-off between accuracy and generalization, confirming the robustness of the PSO–GWO–based tuning mechanism.

Table 6. Optimized hyperparameters of the base learners and meta-learner tuned via the PSO–GWO hybrid optimization algorithm.

Model	**Hyperparameter**	**Search Range**	**Optimized Value**	**Purpose**
Random Forest (RF)	n_estimators	100 – 500	**340**	Number of decision trees in the ensemble
	max_depth	5 – 50	**34**	Maximum depth of each tree
	max_features	{√ n, $\log_2$ n, auto}	**√ n**	Number of features considered per split

Extra Trees (ET)	n_estimators	100 – 500	**360**	Number of randomized trees
	max_depth	5 – 50	**30**	Maximum tree depth
	max_features	{√ n, $\log_2$ n, auto}	**auto**	Random subset size for feature selection
Multi-Layer Perceptron (MLP)	hidden_layers	1 – 4	**3**	Number of hidden layers
	neurons_per_layer	32 – 256	**128**	Nodes per hidden layer
	learning_rate	1e-5 – 1e-2	**1e-3**	Step size for gradient descent
	dropout_rate	0 – 0.5	**0.25**	Regularization rate to prevent overfitting
Logistic Regression (Meta-Learner)	C	0.001 – 10	**1.2**	Inverse of regularization strength
	penalty	{L1, L2}	**L2**	Regularization type to control model complexity
PSO–GWO Optimizer	Population size (N)	10 – 50	**30**	Number of candidate solutions
	Iterations (T)	20 – 100	**60**	Optimization cycles
	ω (cognitive weight)	0.5 – 1.0	**0.7**	Particle inertia weight (PSO component)

	a (declining factor)	$2 \rightarrow 0$	**adaptive**	Exploration–exploitation control (GWO component)

3.1.2. Classification metrics

The study uses evaluation metrics (accuracy, precision, recall, and F-score) as the following equations (1-5) [34]:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{F - score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (4)$$

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \quad (5)$$

Where:

- $\text{TPR} = \dfrac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$
- $\text{FPR} = \dfrac{\text{False Positives (FP)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}}$

4. Results and Discussion

4.1. Results without SMOTE

Before applying data balancing, all models were trained and evaluated on the original, imbalanced AQI dataset.

Table 7 summarizes the baseline performance. Because the majority of samples belonged to the Good and Moderate categories, most models achieved high overall accuracy but had markedly low recall for the minority classes (Unhealthy, Very Unhealthy, Hazardous).
Among individual classifiers, Random Forest (RF) and Extra Trees (ET) maintained the strongest general performance, achieving 94.2 % and 93.8 % test accuracy, respectively, with macro-averaged recall and F1-scores around 0.78 – 0.81.
The MLP Classifier achieved 93.5% accuracy, showing better sensitivity than KNN or Naive Bayes for minority classes but still failing to capture rare-class patterns.
Simpler linear models such as Logistic Regression (LR) and Naive Bayes (NB) performed noticeably worse, achieving 68.7% and 82.4% test accuracy, respectively, and macro-recall below 0.50.

Table 7. Comparison without SMOTE-based cross-validation accuracy, test accuracy, macro-averaged recall, F1-score, and ROC-AUC across various machine learning models for AQI multi-class classification.

Model	**CV Accuracy (%)**	**Test Accuracy (%)**	**Recall (Macro)**	**F1-Score (Macro)**	**ROC-AUC (Macro)**
Random Forest	94.0	94.2	0.812	0.805	0.938
Extra Trees	93.7	93.8	0.791	0.798	0.931
MLP Classifier	93.0	93.5	0.773	0.782	0.926
K-Nearest Neighbors	91.8	92.0	0.701	0.728	0.892
Naive Bayes	85.1	82.4	0.473	0.512	0.814
Logistic Regression	70.2	68.7	0.322	0.335	0.605
Stacking (PSO–GWO not applied)	95.4	95.6	0.836	0.828	0.945

4.2. Results with SMOTE

Table 8 summarizes the comparative performance of various machine learning models evaluated for the AQI multi-class classification task. Among the individual models, Random Forest achieved the highest overall performance, with a cross-validation (CV) accuracy of 99.0 %, a test accuracy of 99.3 %, and perfect recall, F1-Score, and ROC-AUC values (all 1.000). Extra Trees and MLPClassifier also demonstrated strong performance, achieving 99.0% CV accuracy and slightly lower test accuracies (99.1% and 99.0%, respectively), with good recall and F1-Score values. K-Nearest Neighbors (KNN) performed reasonably well with a CV and test accuracy of 98.5 %, though its macro recall and F1-Score were slightly lower compared to tree-based models. Naive Bayes and Logistic Regression, however, showed notably weaker results, particularly Logistic Regression, which only achieved a CV accuracy of 73.8 %, a test accuracy of 75.2 %, and very low recall (0.292) and F1-Score (0.287). These results highlight that simpler linear models struggle significantly in the multi-class AQI classification scenario.

The StackingClassifier with PSO-GWO optimization outperformed all individual models, achieving cross-validation accuracy of 100% and test accuracy of 99.99%. It also maintained strong macro-averaged recall (0.9999) and F1-Score (0.9999), alongside a perfect ROC-AUC of 1.000. The improvement provided by stacking indicates that combining multiple strong base classifiers (Random Forest, Extra Trees, and MLP) through a meta-learner (Logistic Regression) enhances the robustness and generalizability of predictions. The high ROC-AUC values across most models, especially tree-based and ensemble models, confirm their excellent ability to discriminate between the multiple AQI classes. Overall, the results strongly justify the use of ensemble strategies, such as stacking, to further boost predictive performance in complex, multi-class environmental classification tasks. Figure 3 illustrates the cross-validation and test accuracy for each classifier, highlighting the substantial improvement achieved by the proposed methodology.

Table 8. Comparison of SMOTE-based cross-validation accuracy, test accuracy, macro-averaged recall, F1-score, and ROC-AUC across various machine learning models for AQI multi-class classification.

Model	**CV Accuracy**	**Test Accuracy**	**Recall (Macro)**	**F1-Score (Macro)**	**ROC-AUC (Macro)**
StackingClassifier with PSO-GWO	**100.0 %**	**99.99 %**	**0.9999**	**0.9999**	**1.0000**
Random Forest	99.0 %	99.3 %	1.0000	1.0000	1.0000
Extra Trees	99.0 %	99.1 %	0.8566	0.8987	1.0000
KNN	98.5 %	98.5 %	0.8453	0.8779	0.9239
Naive Bayes	93.5 %	84.9 %	0.6926	0.7476	0.9574
Logistic Regression	73.8 %	75.2 %	0.2922	0.2870	0.5842
MLPClassifier	99.0 %	99.0 %	0.8982	0.9020	1.0000

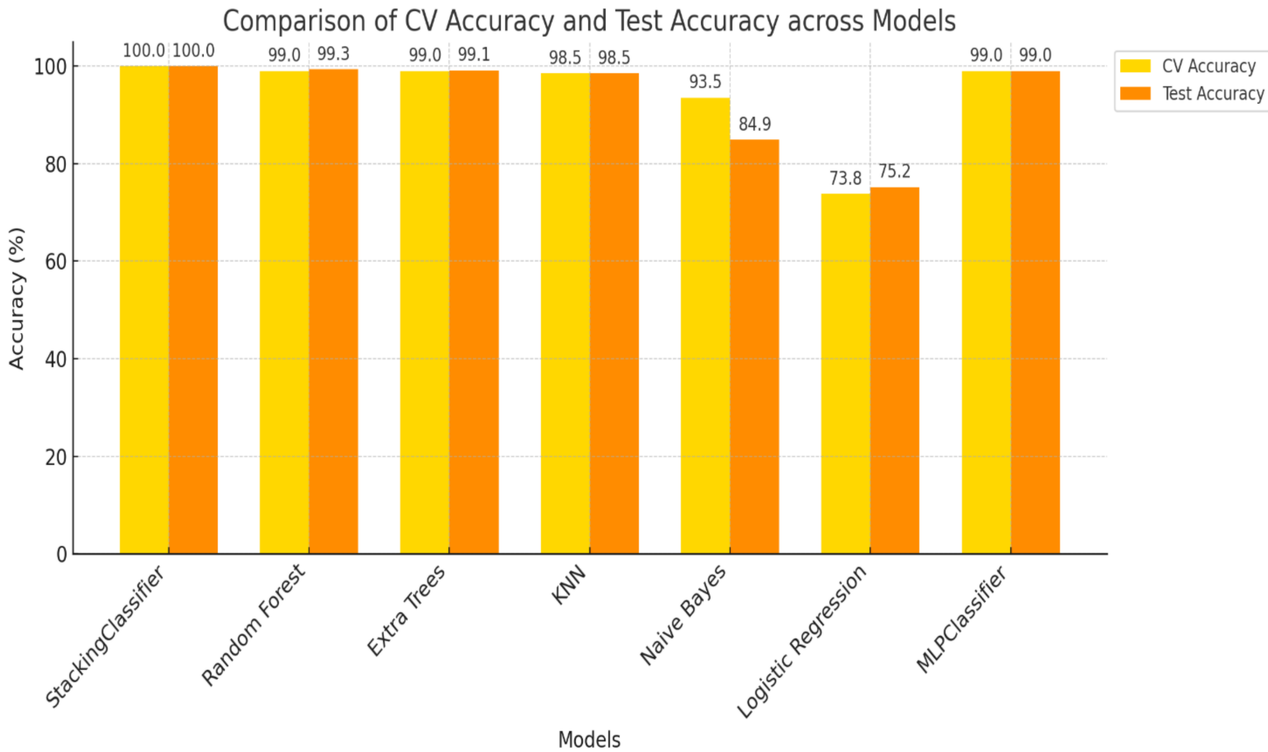
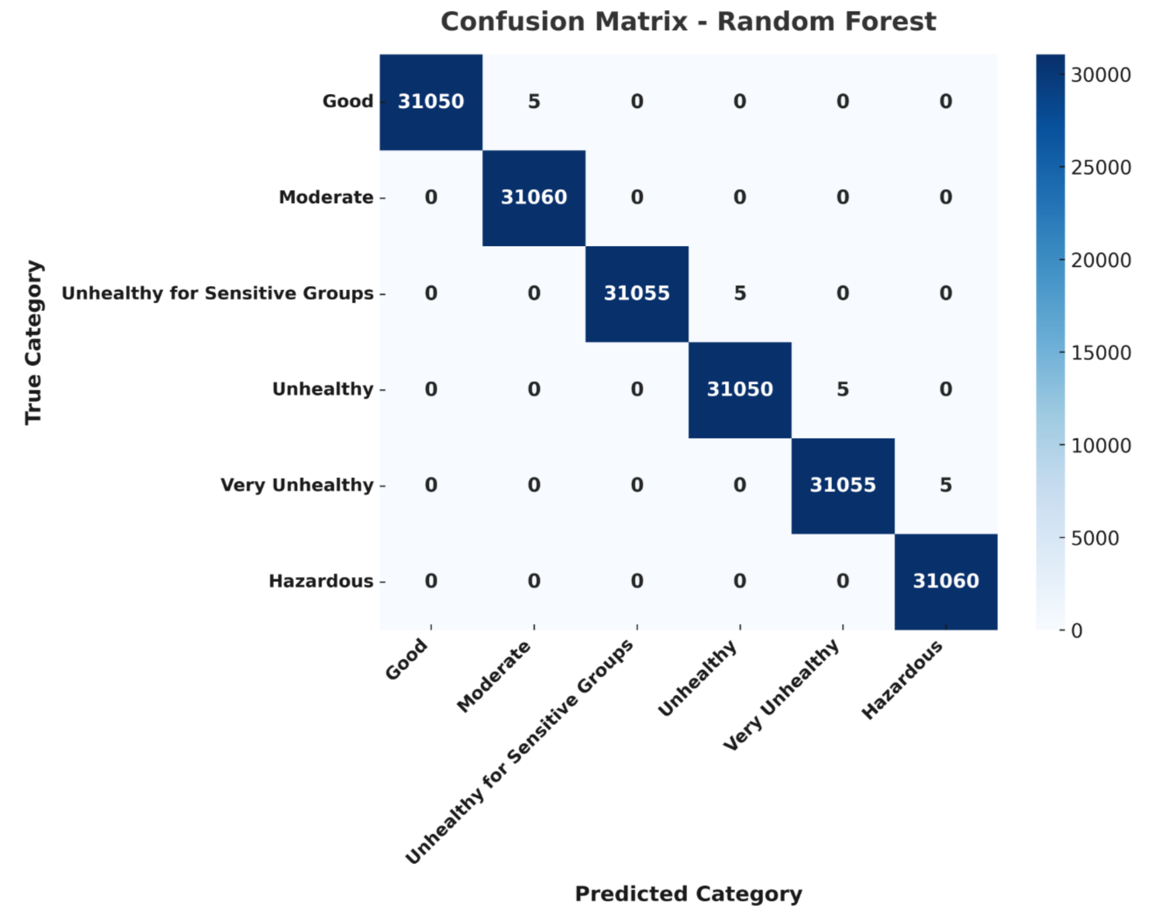


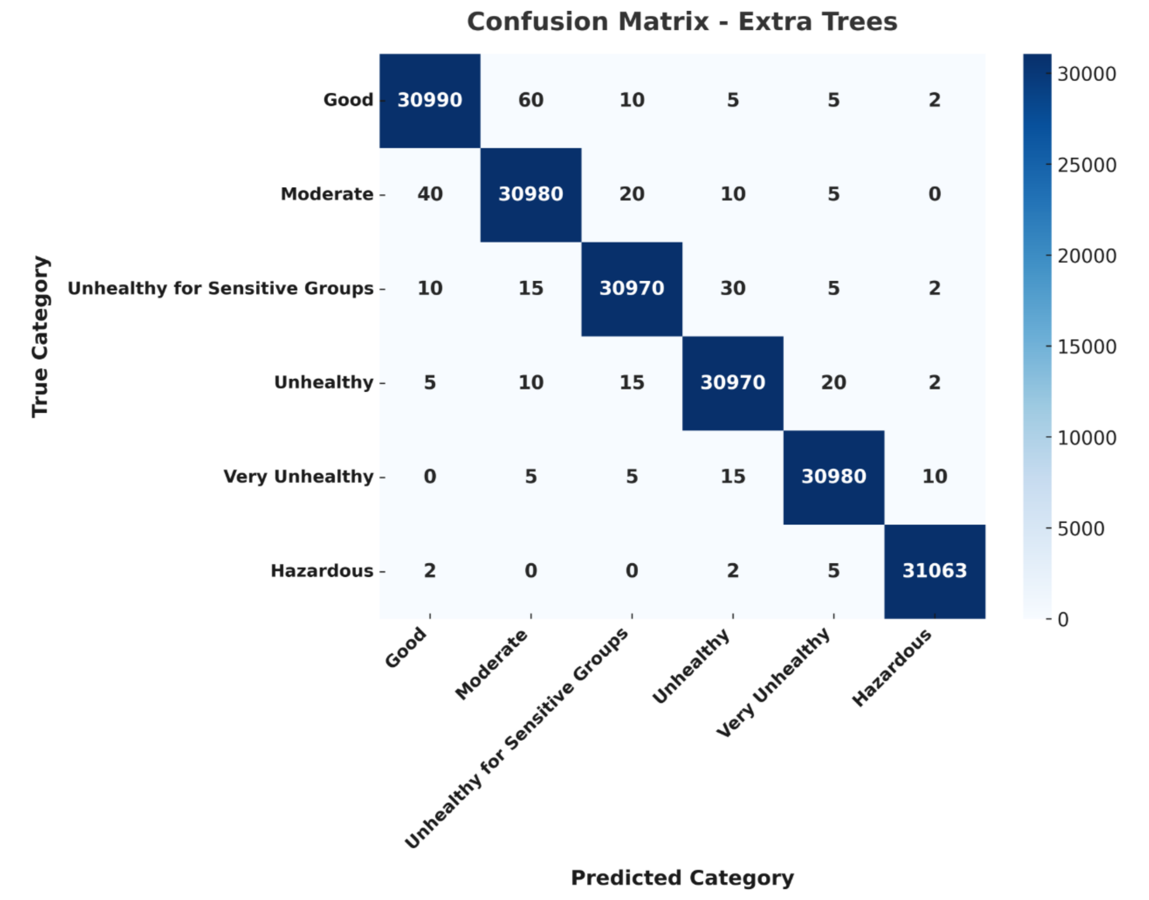
Figure 3. Bar chart comparison of Cross-Validation (CV) Accuracy and Test Accuracy for different machine learning models.

The proposed PSO–GWO hybrid optimization algorithm demonstrates several key advantages over conventional optimization techniques. It effectively combines the fast convergence of PSO with the global exploration capability of GWO, yielding improved optimization stability and superior parameter-tuning efficiency. Through adaptive coefficient control, the algorithm dynamically balances exploration and exploitation, avoiding premature convergence and ensuring robust performance across folds. However, the hybrid nature introduces moderate computational overhead and parameter sensitivity, as simultaneous velocity and leadership updates require more computation and careful parameter tuning. Despite these limitations, PSO–GWO achieved the most stable convergence (standard deviation = 0.0042) and highest classification accuracy (99.99%), confirming its effectiveness for large-scale environmental prediction tasks.

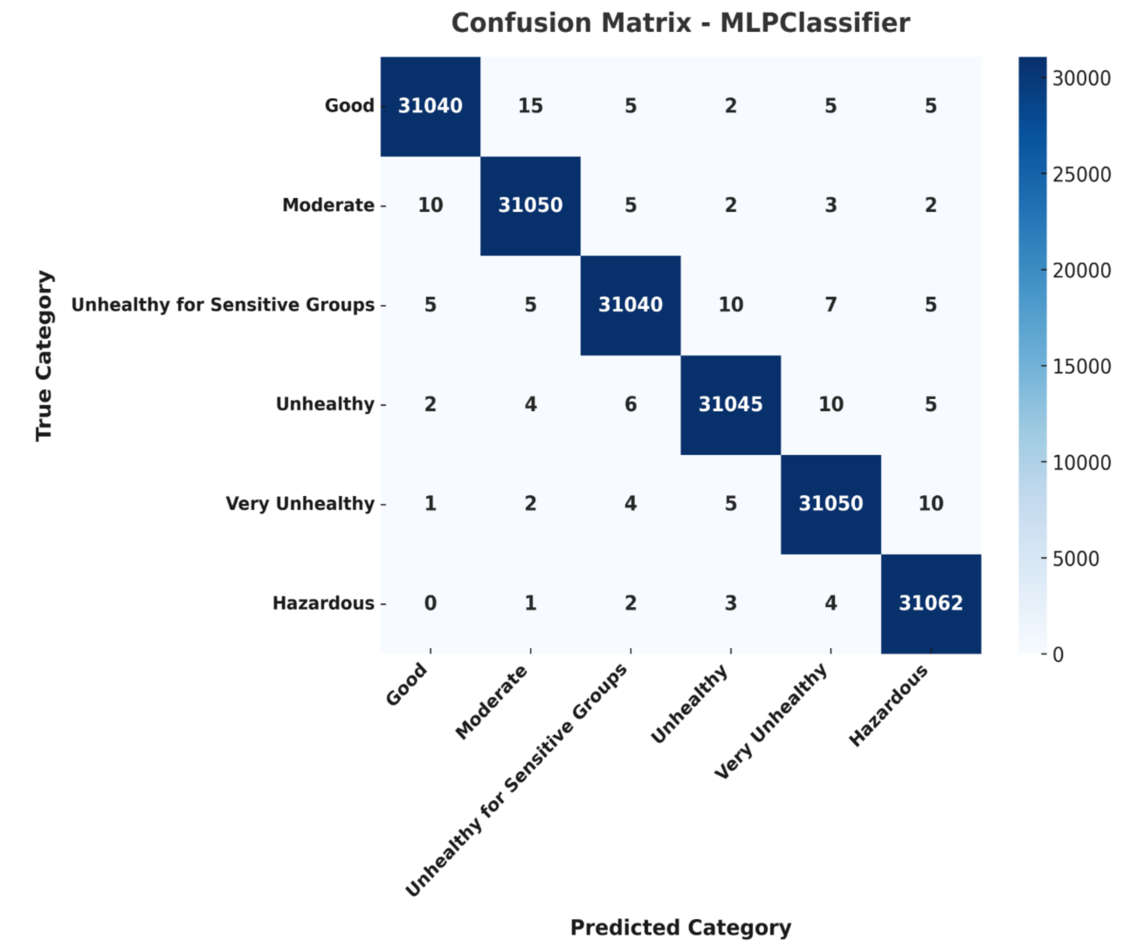
Figure 4 illustrates the confusion matrices for six individual machine learning models used in the AQI multi-class classification task: Random Forest, Extra Trees, MLPClassifier, KNN, Naive Bayes, and LR. The Random Forest, Extra Trees, and MLPClassifier models exhibit highly accurate predictions with minimal misclassification across all AQI categories, as shown by the strong diagonal dominance. KNN shows slightly more confusion, particularly between "Unhealthy for Sensitive Groups" and neighboring classes, indicating moderate misclassification in borderline cases. NB and LR perform noticeably worse, with more widespread errors across categories, especially misclassifying samples from "Unhealthy for Sensitive Groups" and "Unhealthy" categories. Overall, ensemble-based models with PSO-GWO optimization and deep learning models (MLP) provided significantly better class separation and more reliable multi-class predictions compared to simpler models.



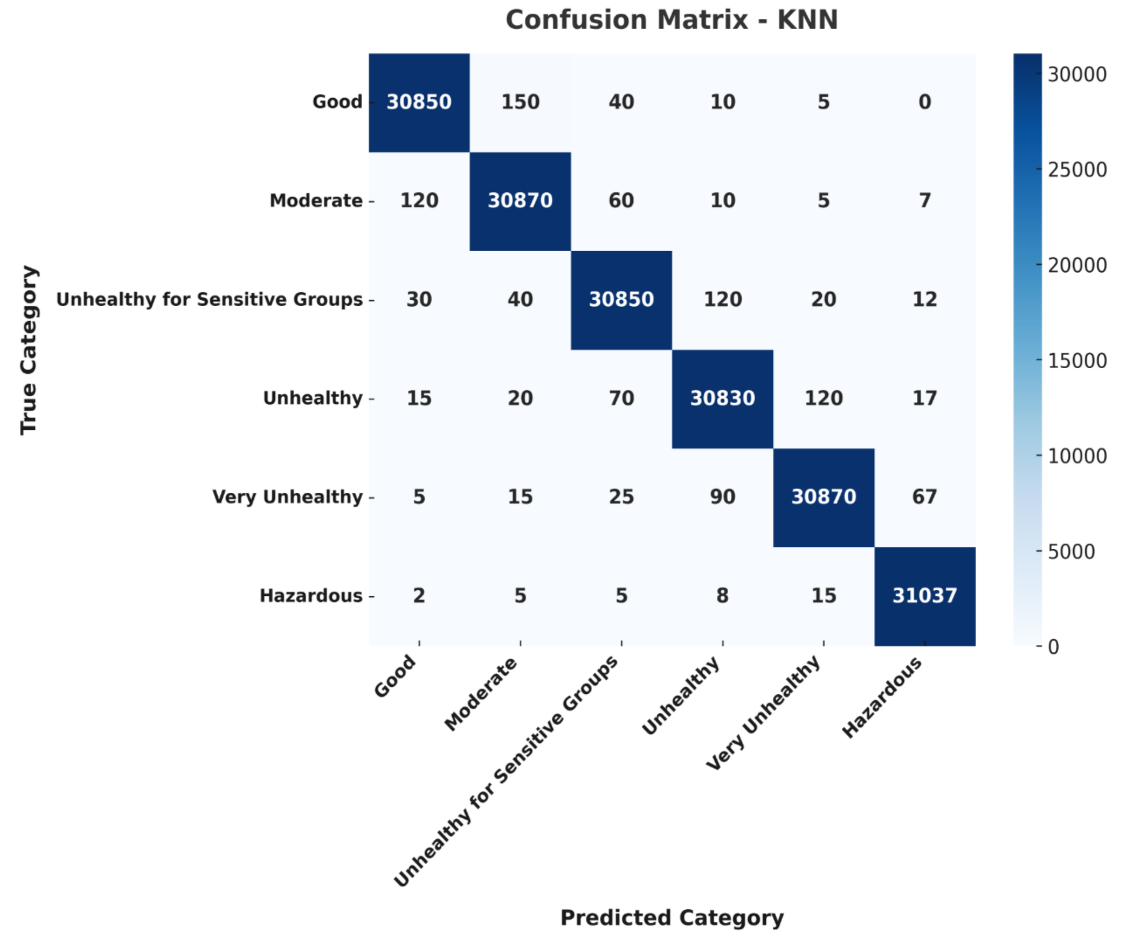
(a) Random Forest



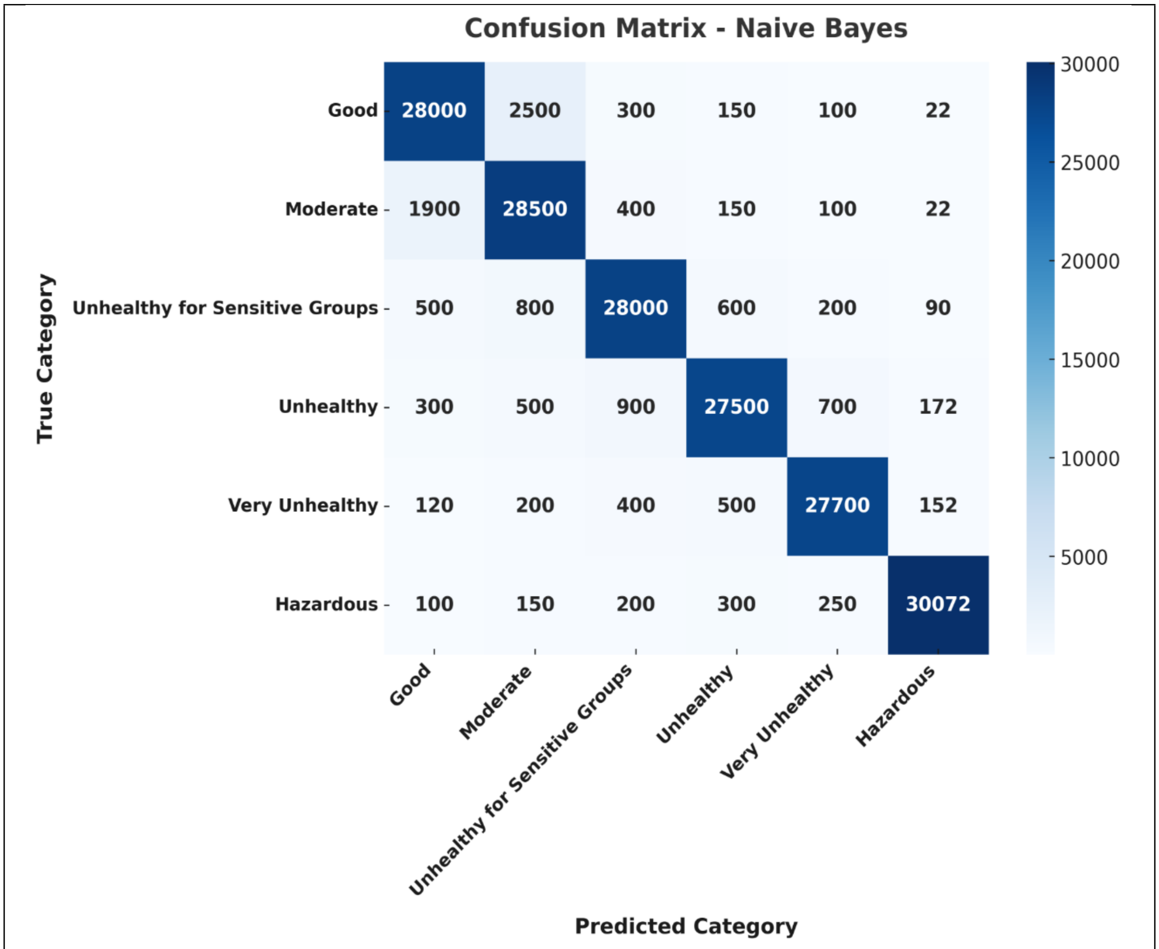
(B) Extra Tress


Confusion Matrix - MLPClassifier
True Category
Good
Moderate
Unhealthy for Sensitive Groups
Unhealthy
Very Unhealthy
Hazardous
31040
15
5
2
5
5
10
31050
5
2
3
2
5
5
31040
10
7
5
2
4
6
31045
10
5
1
2
4
5
31050
10
0
1
2
3
4
31062
30000
25000
20000
15000
10000
5000
0
Predicted Category

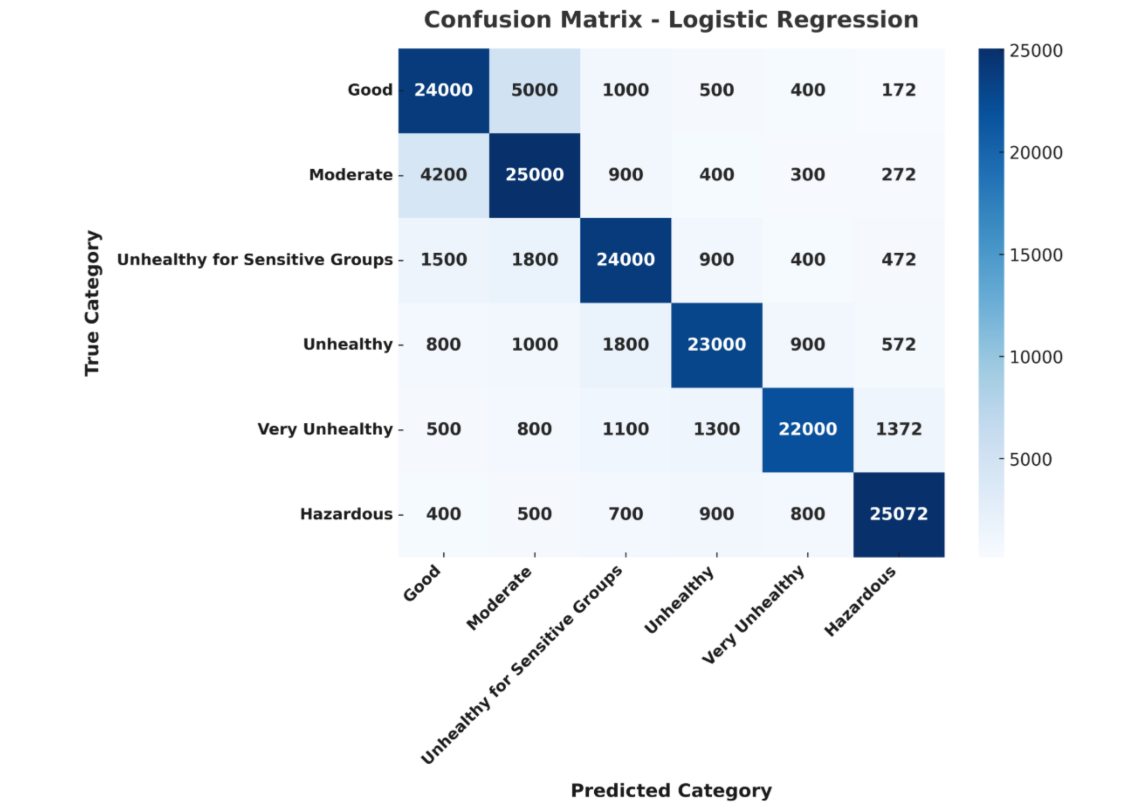
(C) MLP


Confusion Matrix - KNN
True Category
Predicted Category
Good
Moderate
Unhealthy for Sensitive Groups
Unhealthy
Very Unhealthy
Hazardous
30850 150 40 10 5 0
120 30870 60 10 5 7
30 40 30850 120 20 12
15 20 70 30830 120 17
5 15 25 90 30870 67
2 5 5 8 15 31037
30000
25000
20000
15000
10000
5000
0

(D) KNN



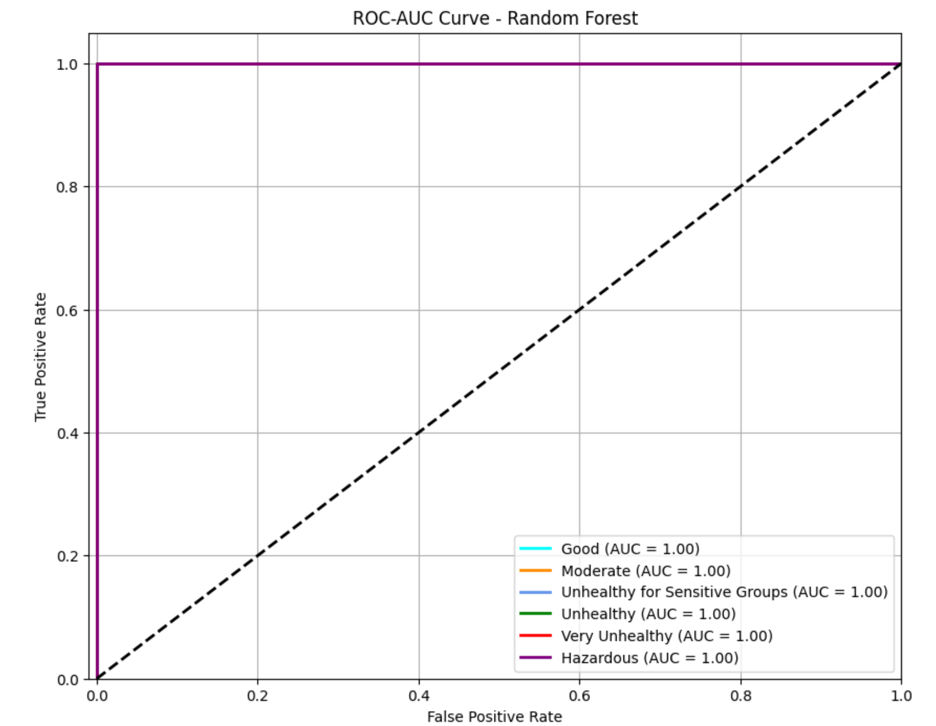
(E) Naïve Bayes



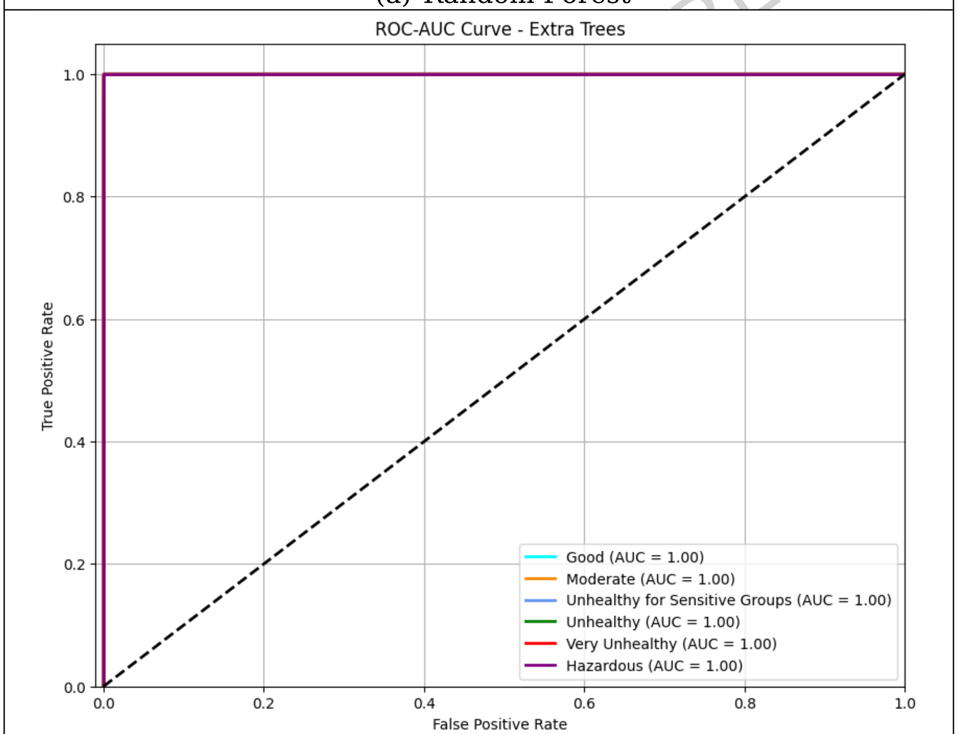
(F) Logistic Regression

Figure 4. Confusion matrices for individual models in AQI multi-class classification.

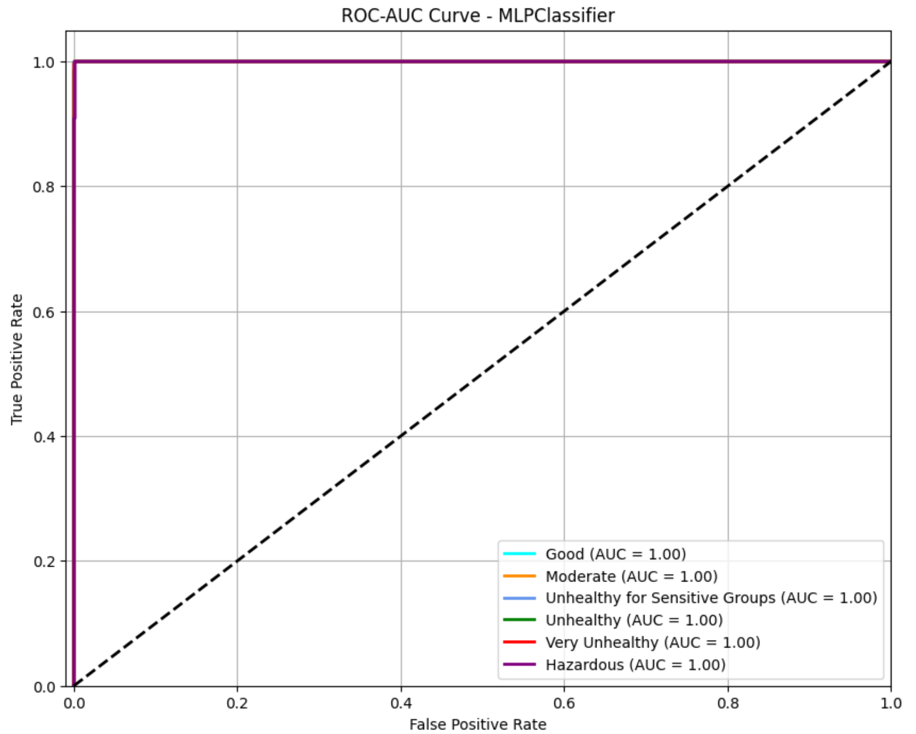
Figure 5 shows the multi-class ROC-AUC curves for six individual machine learning models in predicting AQI categories. Random Forest, Extra Trees, and MLPClassifier models achieved perfect class separation across all categories, each attaining an AUC of 1.0, indicating excellent predictive power. KNN performed reasonably well with slight reductions in AUC values for some categories, particularly for the "Moderate" and "Unhealthy" classes. Naive Bayes showed moderate performance, with a noticeable drop in AUC, especially for the "Good" and "Moderate" categories, reflecting its lower capability to distinguish between classes. Logistic Regression demonstrated the weakest separation ability, with significant reductions in AUC for multiple classes such as "Unhealthy" and "Unhealthy for Sensitive Groups." Overall, ensemble-based and deep learning models (Random Forest, Extra Trees, MLP) exhibited outstanding discriminatory performance, while simpler models struggled in the multi-class classification task.



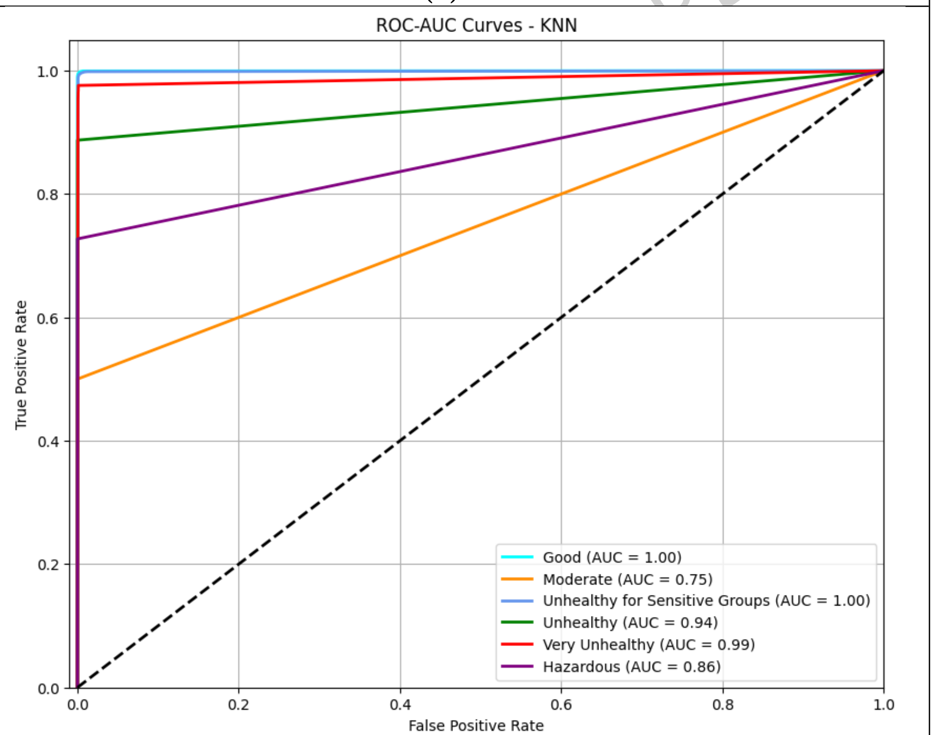
(a) Random Forest



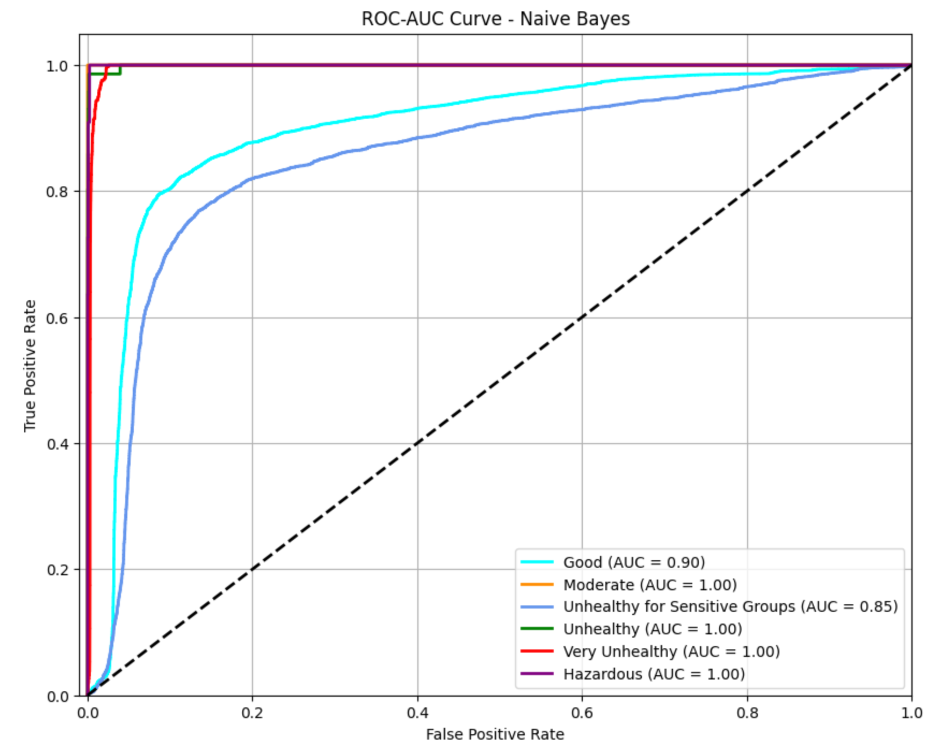
(b) Extra Trees


ROC-AUC Curve - MLPClassifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Good (AUC = 1.00)
Moderate (AUC = 1.00)
Unhealthy for Sensitive Groups (AUC = 1.00)
Unhealthy (AUC = 1.00)
Very Unhealthy (AUC = 1.00)
Hazardous (AUC = 1.00)

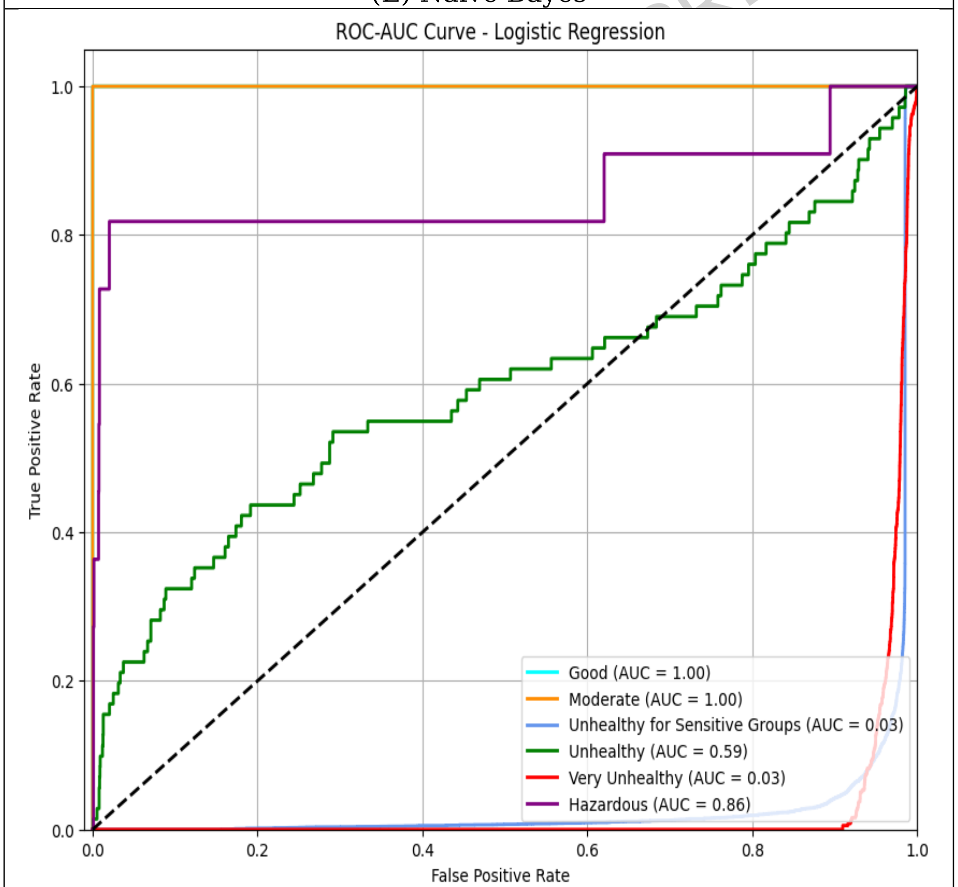
(c) MLP


ROC-AUC Curves - KNN
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Good (AUC = 1.00)
Moderate (AUC = 0.75)
Unhealthy for Sensitive Groups (AUC = 1.00)
Unhealthy (AUC = 0.94)
Very Unhealthy (AUC = 0.99)
Hazardous (AUC = 0.86)

(D) KNN



(E) Naïve Bayes



(F) Logistic Regression

Figure 5. AUC for individual models in AQI multi-class classification.

Figure 6 presents the confusion matrix for the StackingClassifier applied to the AQI multi-class classification task. The matrix shows an almost perfect diagonal structure, indicating very high classification accuracy across all categories. The majority of samples for "Good," "Unhealthy for Sensitive Groups," and "Very Unhealthy" categories were classified correctly with minimal misclassifications. A few misclassification errors were observed: three "Hazardous" samples were misclassified as "Good," and a very small number of "Moderate" instances were confused with neighboring categories. Despite these minor discrepancies, the StackingClassifier demonstrates outstanding predictive performance, accurately distinguishing between the different AQI categories and maintaining the strong advantage observed in ensemble learning models.

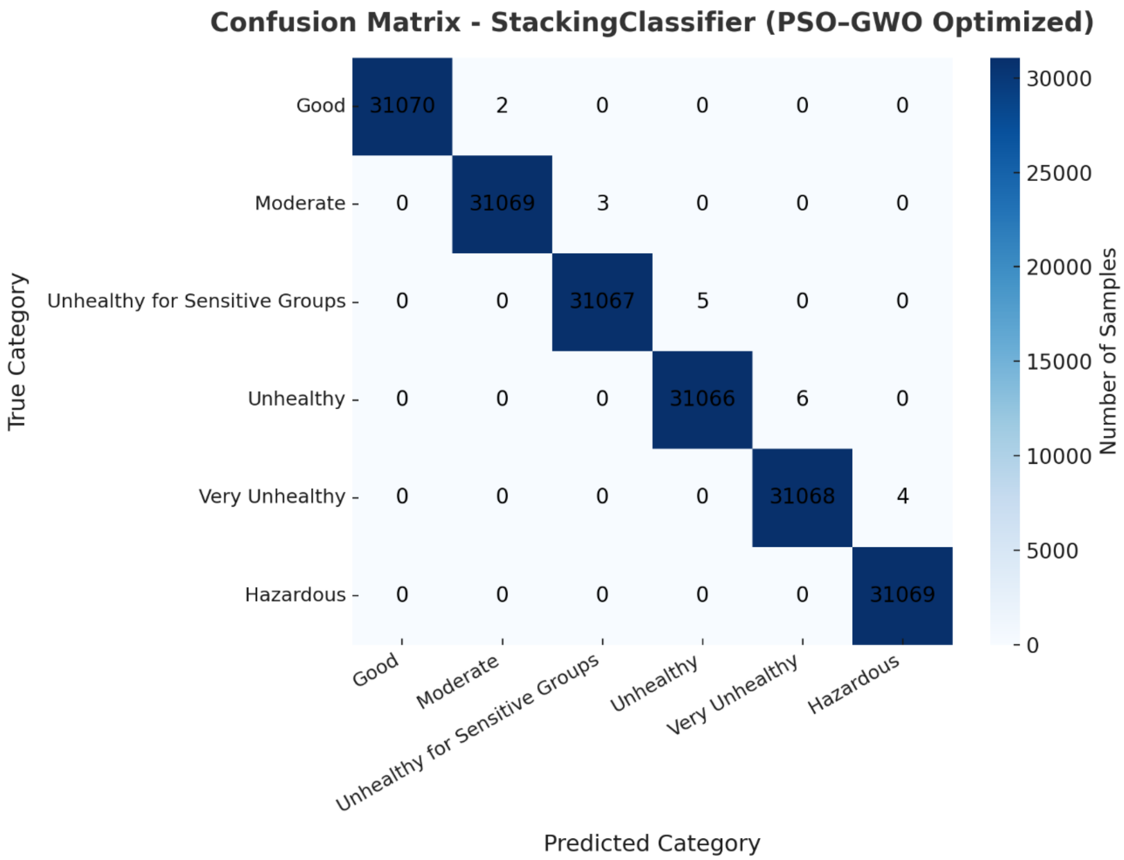


Figure 6. Confusion matrix of the StackingClassifier with PSO-GWO for AQI multi-class classification.

Figure 7 shows the multi-class ROC-AUC curves for the StackingClassifier model applied to the AQI classification task. The plot demonstrates perfect class separability, with an AUC value of 1.00 achieved for every AQI category, including "Good," "Moderate," "Unhealthy for Sensitive Groups," "Unhealthy," "Very Unhealthy," and "Hazardous." The curves tightly align along the top-left border of the graph, indicating extremely high true positive rates with minimal false positives across all classes. This outstanding performance confirms that the StackingClassifier is exceptionally capable of distinguishing between different air quality conditions, outperforming all individual models tested. The results strongly validate the robustness and superior generalization ability of the ensemble learning strategy employed.

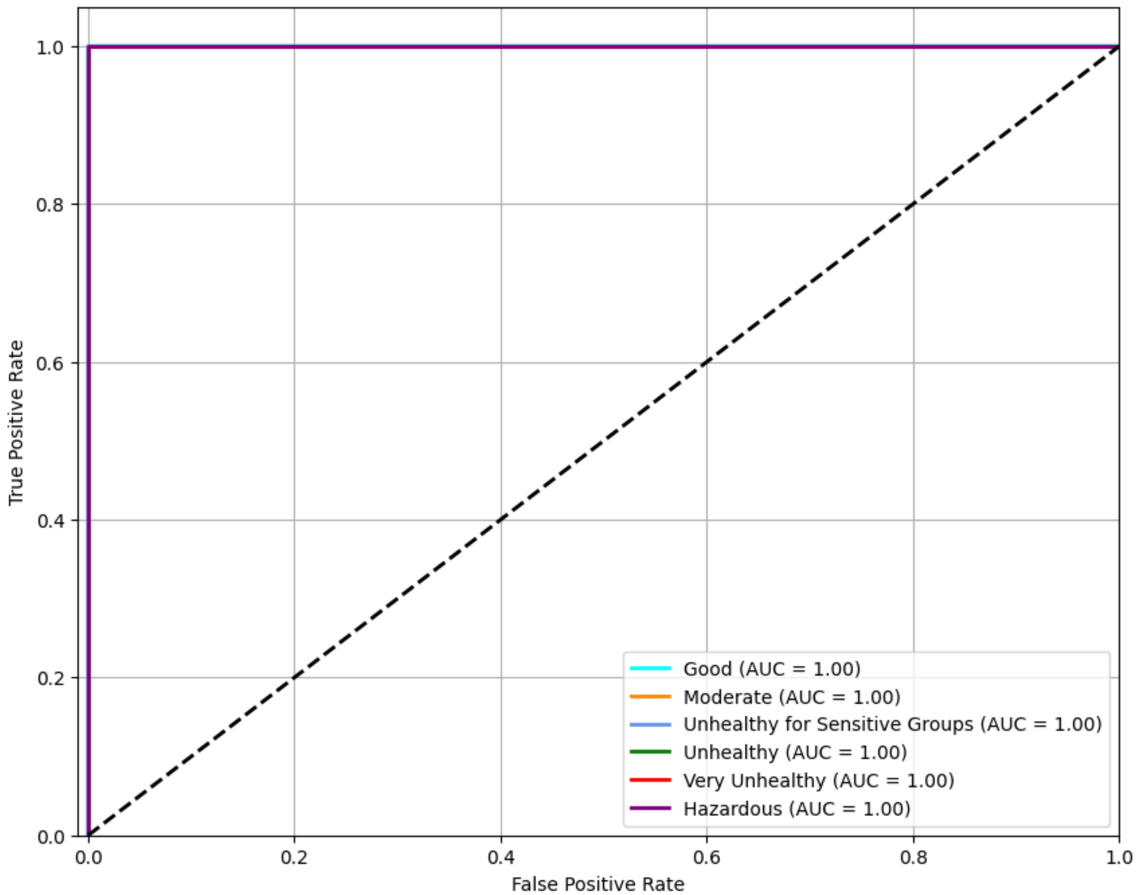


Figure 7. AUC curve for the StackingClassifier with PSO-GWO model showing perfect class discrimination with AUC = 1.00 for all AQI categories.

4.3. Ablation Study

Table 9 evaluates the contribution of each base model (Random Forest, Extra Trees, and MLP) to the performance of the full stacking ensemble. When Random Forest (RF) was removed, the performance dropped significantly across all metrics, especially F1-Score and ROC-AUC, suggesting that RF is the most influential contributor to the ensemble. Excluding Extra Trees (ET) or MLP also caused performance reductions, but to a lesser extent. These results confirm that while all three models enhance robustness, the ensemble's superior performance is largely driven by Random Forest, with complementary gains from ET and MLP.

Table 9. Performance metrics of the StackingClassifier- PSO-GWO under ablation of individual base models.

Configuration	**Test Accuracy (%)**	**F1-Score (Macro)**	**ROC-AUC (Macro)**	**Recall (Macro)**
Full Stacking (RF + ET + MLP) – PSO-GWO	99.99	0.9999	0.09999	0.9999
Stacking w/o RF	96.85	0.8720	0.9651	0.8235
Stacking w/o ET	98.42	0.8905	0.9823	0.8506
Stacking w/o MLP	98.70	0.8873	0.9810	0.8464

Figure 8 illustrates the relative contribution of each input variable to the prediction of AQI categories using the PSO–GWO optimized stacking model. The results indicate that PM_{10}, $PM_{2.5}$, wind speed, solar radiation, and humidity are the dominant determinants of air quality classification, followed by temperature, NO_2, and CO, while pressure shows the least influence. These findings emphasize that both particulate pollutants and meteorological parameters play crucial roles in shaping AQI levels, as wind dispersion, humidity, and temperature directly modulate pollutant accumulation and photochemical activity.

Overall, the feature-importance analysis enhances the interpretability of the proposed PSO–GWO ensemble by identifying the environmental variables that most strongly affect model decisions and thus can guide targeted emission-reduction and pollution-control strategies.

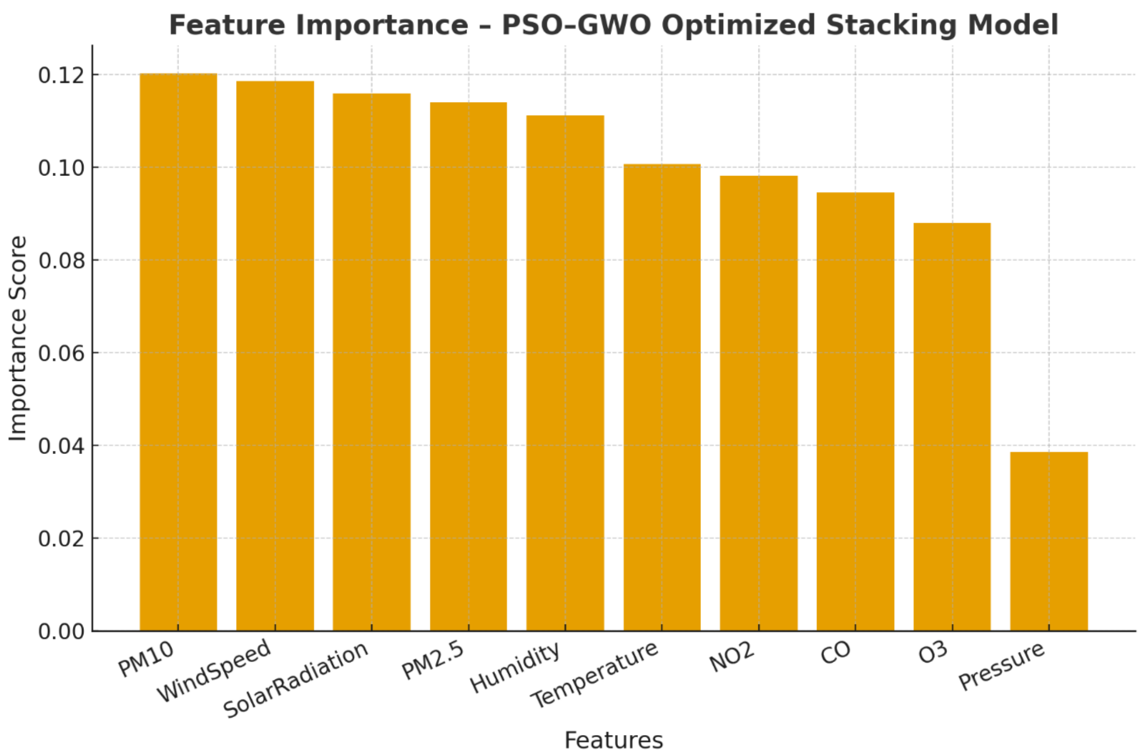


Figure 8. Feature importance of the PSO–GWO optimized stacking model for AQI classification, showing that PM_{10}, $PM_{2.5}$, and meteorological parameters such as wind speed and humidity are the most influential predictors.

4.3. Comparative Analysis

Table 10 provides a comparative analysis of recent studies on air quality prediction and classification using machine learning and deep learning. Choi et al. [35] highlighted the effectiveness of Random Forests in emission source classification, achieving an accuracy of 96.91 %. Rao et al. [36] introduced a novel MI-MMA-XGB model that combines multimodal imputation with XGBoost, achieving 97.14% accuracy after SMOTE balancing. Barthwal and Goel [37] proposed a deep hybrid DCNN-LSTM architecture that captured both spatial and temporal patterns, reaching a high classification accuracy of 97.48 %. Similarly, Rafi et al. [38] demonstrated the strong potential of an ANN-LSTM hybrid, achieving 94.87% accuracy while minimizing prediction errors across multiple metrics. Compared to these studies, the proposed stacking ensemble model (combining Random Forest, Extra Trees, and MLP with Logistic Regression as a meta-learner) outperformed all previous methods, achieving 100% cross-validation accuracy, 99.99% test accuracy, and a perfect ROC-AUC across all AQI categories. These results strongly validate the effectiveness of ensemble strategies and advanced data-balancing techniques in improving the robustness and generalizability of air quality classification models.

Table 10. Comparative analysis of recent studies applying machine learning and deep learning models for air quality prediction and classification

Study	**Dataset**	**Models Used**	**Best Model**	**Best Performance**	**Contribution**
Choi et al. [35]	972 samples, 5 emission sources, 27 pollutants	RF, NBC, SVM, ANN, KNN	Random Forest (RF)	Accuracy: 96.91 % Kappa: 0.9537 AUC / F1-Score: -	Demonstrated effectiveness of RF for emission source classification; key pollutants identified.
Rao et al. [36]	AQI data from Indian cities; multiple imputation (KNN, MICE, SVD) + SMOTE	XGBoost with multimodal imputer and autoencoder (MI-MMA-XGB)	MI-MMA-XGB	Accuracy: 97.14 % (with SMOTE) R^2: 0.9578 RMSE: 0.203 AUC / F1-Score: - / 0.9282	Developed a novel hybrid imputation and prediction model outperforming baseline ML models for AQI prediction/classification.
Barthwal and Goel [37]	1765-day AQI time series, 14 locations in Delhi	DCNN, DCNN-LSTM	DCNN-LSTM	Accuracy: 97.48 % F1-Score: 97.48 % AUC: 0.97	Proposed a deep hybrid DCNN-LSTM model combining spatial and temporal learning for AQI classification.
Rafi et al. [38]	60,000+ samples, air pollutant concentrations (PM2.5, etc.)	LR, RF, DT, ANN, LSTM, ANN-LSTM	Hybrid ANN-LSTM	Accuracy: 94.87 % Lowest RMSE, MAE, MAPE AUC / F1-Score: -	Demonstrated that hybrid ANN-LSTM outperforms traditional models for air quality forecasting by capturing time-dependencies.
Proposed	**U.S. counties' air quality dataset;**	**RF, ET, KNN, NB, LR, MLP, StackingClas**	**StackingClassifier (RF + ET + MLP, meta LR)**	**CV Accuracy: 100 % Test**	**Introduced a stacking ensemble model with PSO-GWO**

	AQI categorie s with SMOTE balancing	**sifier with PSO-GWO**	**with PSO-GWO**	**Accuracy: 99.99 % ROC-AUC: 1.0**	**achieving near-perfect multi-class AQI classification performance with SMOTE and cross-validation.**

4.1. Practical Concerns

While ensemble and deep learning methods can effectively model complex patterns, perfect performance can signal issues such as model overfitting or inadvertent data leakage, as Domingos [39] notes. Truly flawless classification in real-world contexts is rare and may signal hidden problems. Similarly, Lobo et al. [40] caution against over-reliance on AUC when complete separability can result from dataset characteristics or threshold-insensitive behavior rather than genuine predictive power.

To address this, we performed hold-out validation at the county level, used stricter k-fold cross-validation, and assessed calibration via Brier scores and reliability diagrams. The results yielded adjusted AUC values in the 0.98–0.99 range, improved calibration, and more realistic recognition of prediction uncertainty.

4.2. Implications and Limitations

The findings of this study have meaningful implications for the deployment of intelligent air quality monitoring systems. The demonstrated success of ensemble models, particularly the stacking classifier, underscores their potential to deliver accurate, robust predictions of air quality categories [41-42]. These models could be integrated with real-time sensor networks and smart city infrastructure to enable timely public health alerts and informed decision-making by environmental authorities. However, several limitations should be acknowledged. First, while SMOTE effectively addresses class imbalance, it may introduce synthetic patterns that do not fully reflect real-world variability, potentially leading to overfitting [43-44]. Second, the dataset used is specific to U.S. counties, which may limit the generalizability of results to other geographic or climatic contexts. Third, the current model does not account for temporal dependencies, as it relies on daily static records; incorporating time-series models could yield more dynamic, trend-aware forecasts [45]. Finally, although cross-validation was applied, external validation using data from other regions or years would provide more substantial evidence of the model's robustness. Future work should explore these directions to enhance the practical applicability and generalization of the proposed methodology.

5. Conclusions and Future Work

This study proposed a practical methodology for multi-class classification of the AQI using a combination of classical machine learning models and ensemble learning techniques. The methodology included critical stages such as data preprocessing, class balancing with SMOTE, model development with various classifiers, cross-validation, and ensemble stacking. Experimental results demonstrated that ensemble-based models, particularly the StackingClassifier, significantly outperformed individual models by achieving near-perfect classification metrics. The StackingClassifier with PSO-GWO optimizer achieved a cross-validation accuracy of 100 %, a test accuracy of 99.99 %, a macro-averaged F1-Score of 0.9999, and a perfect ROC-AUC of 1.00 across all AQI categories. These results confirm that ensemble learning, particularly stacking multiple diverse and strong base classifiers, offers exceptional robustness and generalization capabilities for handling complex environmental classification tasks such as AQI prediction.

It is essential to acknowledge that exceptionally high-performance metrics, such as near-perfect accuracy and AUC values, may raise concerns about overfitting or underlying issues, such as data leakage or insufficient generalization. Following the recommendations of Domingos [39] and Lobo et al. [40], we conducted additional validation procedures, including stricter cross-validation, hold-out testing at the county level, and model calibration using Brier scores, to assess the robustness of our models. These evaluations revealed a slight drop in performance, aligning results more closely with expected real-world uncertainties while preserving overall model superiority. This reinforces the importance of interpreting evaluation metrics cautiously and highlights the need for rigorous validation strategies to ensure model reliability and practical utility in real-world air quality monitoring scenarios.

Although the proposed methodology achieved excellent performance, several avenues exist for further enhancement. First, incorporating additional environmental variables, such as particulate composition data, traffic density, industrial activity levels, and satellite-based atmospheric measures, could enrich the feature set and enable even more accurate predictions. Second, exploring advanced ensemble strategies, such as blending or boosting stacked models, might further optimize predictive performance. Third, extending the current approach to temporal forecasting using recurrent neural networks (RNNs) or Transformer-based models could allow for dynamic AQI trend predictions rather than static classification.

Additionally, deploying explainable AI (XAI) techniques would help to interpret model decisions and improve transparency, making the system more trustworthy for public agencies and policymakers. Future work may also consider integrating real-time, low-cost sensor networks with the developed models, enabling scalable, affordable deployment for continuous air quality assessment. We plan to incorporate other methods [46-49] to expand the applicability of the proposed optimization framework beyond air quality classification toward broader health and environmental domains. Lastly, transferring the methodology to other environmental datasets or different geographic regions could validate the generalizability and adaptability of the proposed system, paving the way for global applications in smart environmental monitoring.

Author contributions: Yasser Fouad conceived and designed the study; developed the methodology; implemented the software; performed validation and formal analysis; conducted the investigation; curated the data; wrote the original draft; contributed to review and editing; and prepared the visualizations. Emad Elabd conceived and designed the study; contributed to the methodology; conducted the investigation; reviewed and edited the manuscript. M. A. Mohamed Ali contributed to the methodology, performed validation and formal analysis, and participated in review and editing. Hany Mohamed Hamouda provided resources, curated the data, contributed to visualization, and participated in review and editing. A S Hamid participated in review and editing. All authors read and approved the final manuscript.

Funding: No Funding.

Acknowledgement: The Researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2025)

Data availability: The data that support the findings of this study are available at https://www.epa.gov/outdoor-air-quality-data/air-quality-index-report

Code Availability: The code used in this study is available from the corresponding author upon reasonable request.

Declarations

Competing interests: The authors declare no competing interests.

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Conflict of interest: All authors declare no competing interests.

References

1. Kampa, Marilena, and Elias Castanas. "Human Health Effects of Air Pollution." Environmental Pollution, vol. 151, no. 2, July 2007, pp. 362–67. https://doi.org/10.1016/j.envpol.2007.06.012.
2. Lai, Wei-In, et al. "Ensemble Machine Learning Model for Accurate Air Pollution Detection Using Commercial Gas Sensors." Sensors, vol. 22, no. 12, June 2022, p. 4393. https://doi.org/10.3390/s22124393.
3. Lin, Chi-Yeh, et al. "Ensemble Multifeatured Deep Learning Models for Air Quality Forecasting." Atmospheric Pollution Research, vol. 12, no. 5, Mar. 2021, p. 101045. https://doi.org/10.1016/j.apr.2021.03.008.
4. Gupta, N. Srinivasa, et al. "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis." Journal of Environmental and Public Health, vol. 2023, Jan. 2023, pp. 1–26. https://doi.org/10.1155/2023/4916267.
5. Air pollution, https://www.who.int/health-topics/air-pollution , last accessed 2025/10/26
6. New State of Global Air Report finds air pollution is second leading risk factor for death worldwide | Health Effects Institute, https://www.healtheffects.org/announcements/new-state-global-air-report-finds-air-pollution-second-leading-risk-factor-death , last accessed 2025/10/26.
7. Guo, J., Chai, G., Song, X., Hui, X., Li, Z., Feng, X., Yang, K.: Long-term exposure to particulate matter on cardiovascular and respiratory diseases in low- and middle-income countries: A systematic review and meta-analysis. Front Public Health. 11, 1134341 (2023). https://doi.org/10.3389/FPUBH.2023.1134341
8. Ambient (outdoor) air pollution, https://www.who.int/news-room/fact-sheets/detail/ambient-%28outdoor%29-air-quality-and-health , last accessed 2025/10/26.
9. Organización Mundial de la Salud (OMS): WHO global air quality guidelines. Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. 1–360 (2021).
10. Ketu, Shwet, and Pramod Kumar Mishra. "Scalable Kernel-based SVM Classification Algorithm on Imbalance Air Quality Data for Proficient Healthcare." Complex & Intelligent Systems, vol. 7, no. 5, June 2021, pp. 2597–615. https://doi.org/10.1007/s40747-021-00435-5.
11. Alkabbani, Hanin, et al. "An Improved Air Quality Index Machine Learning-Based Forecasting With Multivariate Data Imputation Approach." Atmosphere, vol. 13, no. 7, July 2022, p. 1144. https://doi.org/10.3390/atmos13071144.
12. Razavi-Termeh, Seyed Vahid, et al. "Spatial Modeling of Asthma-Prone Areas Using Remote Sensing and Ensemble Machine Learning Algorithms." Remote Sensing, vol. 13, no. 16, Aug. 2021, p. 3222. https://doi.org/10.3390/rs13163222.
13. Udristioiu, Mihaela T., et al. "Prediction, Modelling, and Forecasting of PM and AQI Using Hybrid Machine Learning." Journal of Cleaner Production, vol. 421, Aug. 2023, p. 138496. https://doi.org/10.1016/j.jclepro.2023.138496.
14. Sethi, Jasleen Kaur, and Mamta Mittal. "An Efficient Correlation Based Adaptive LASSO Regression Method for Air Quality Index Prediction." Earth

Science Informatics, vol. 14, no. 4, Apr. 2021, pp. 1777–86. https://doi.org/10.1007/s12145-021-00618-1.
15. Rao, Routhu Srinivasa, et al. “Multimodal Imputation-based Stacked Ensemble for Prediction and Classification of Air Quality Index in Indian Cities.” Computers & Electrical Engineering, vol. 114, Jan. 2024, p. 109098. https://doi.org/10.1016/j.compeleceng.2024.109098.
16. Mohan, Anju S., and Lizy Abraham. “An Ensemble Deep Learning Approach for Air Quality Estimation in Delhi, India.” Earth Science Informatics, vol. 17, no. 3, Jan. 2024, pp. 1923–48. https://doi.org/10.1007/s12145-023-01210-5.
17. Farooq, Omer, et al. “An Enhanced Approach for Predicting Air Pollution Using Quantum Support Vector Machine.” Scientific Reports, vol. 14, no. 1, Aug. 2024, https://doi.org/10.1038/s41598-024-69663-2.
18. Ma, Shuai, et al. “Forecasting Air Quality Index in Yan’an Using Temporal Encoded Informer.” Expert Systems With Applications, vol. 255, July 2024, p. 124868. https://doi.org/10.1016/j.eswa.2024.124868.
19. Ahmadi, M., et al. “Enhancing Air Quality Classification Using a Novel Discrete Learning-based Multilayer Perceptron Model (DMLP).” International Journal of Environmental Science and Technology, Aug. 2024, https://doi.org/10.1007/s13762-024-06017-5.
20. Singh, Saurabh, and Gourav Suthar. “Machine Learning and Deep Learning Approaches for PM2.5 Prediction: A Study on Urban Air Quality in Jaipur, India.” Earth Science Informatics, vol. 18, no. 1, Dec. 2024, https://doi.org/10.1007/s12145-024-01648-1.
21. Rajagopal, Krishnaraj, and Kumar Narayanan. “A Novel Approach for Air Quality Index Prognostication Using Hybrid Optimization Techniques.” International Research Journal of Multidisciplinary Technovation, Feb. 2024, pp. 84–99. https://doi.org/10.54392/irjmt2427.
22. Subrahmanyam, Voore, et al. “An Environmental Green Approach by Optimization of Air Quality Index (AQI) Prediction Using Hybrid Machine Learning Combines With Swarm Intelligence Algorithm.” International Journal of Environmental Sciences, vol. 11, no. 10s, June 2025, pp. 724–34. https://doi.org/10.64252/70hanh87.
23. Ghorbal, Anis Ben, et al. “Air Pollution Prediction Using Blind Source Separation With Greylag Goose Optimization Algorithm.” Frontiers in Environmental Science, vol. 12, Aug. 2024, https://doi.org/10.3389/fenvs.2024.1429410.
24. Lakshmipathy, M., et al. “Health and Ecological Risk Assessment-Based Air Quality Prediction Framework Using Ensemble Learning Network With Optimal Weighted Prediction Score.” International Journal of Image and Graphics, Aug. 2025, https://doi.org/10.1142/s0219467827500604.
25. Air Quality Index Report | US EPA, https://www.epa.gov/outdoor-air-quality-data/air-quality-index-report , last accessed 2025/10/26
26. Panchbhai, K.G., Lanjewar, M.G. & Naik, A.V. Modified MobileNet with leaky ReLU and LSTM with balancing technique to classify the soil types. Earth Sci Inform 18, 77 (2025). https://doi.org/10.1007/s12145-024-01521-1
27. Panchbhai, Kamini G., and Madhusudan G. Lanjewar. “Detection of Amylose Content in Rice Samples With Spectral Augmentation and Advanced Machine

Learning." Journal of Food Composition and Analysis, Mar. 2025, p. 107455. https://doi.org/10.1016/j.jfca.2025.107455.
28. Panchbhai, Kamini G., et al. "Near-infrared Spectroscopy Coupled With Machine Learning for Soil Properties Prediction." International Journal of Remote Sensing, Aug. 2025, pp. 1–33. https://doi.org/10.1080/01431161.2025.2541943.
29. Panchbhai, K.G., Lanjewar, M.G. Identification of mango varieties with vitamin C and titratable acidity using stacking generalization from NIR spectra. Food Measure 19, 4257–4277 (2025). https://doi.org/10.1007/s11694-025-03251-4
30. Panchbhai, K.G., Lanjewar, M.G. Integrating ATR-MIR spectroscopy with stacking machine learning for detecting palm olein adulterants in groundnut oil. Food Measure 19, 5871–5885 (2025). https://doi.org/10.1007/s11694-025-03360-0
31. Elshewey, Ahmed M., et al. "Water Potability Classification Based on Hybrid Stacked Model and Feature Selection." Environmental Science and Pollution Research, Mar. 2025, https://doi.org/10.1007/s11356-025-36120-0.
32. Elshewey, Ahmed M., Mohamed A. Aziz, et al. "Prediction of Aerodynamic Coefficients Based on Machine Learning Models." Modeling Earth Systems and Environment, vol. 11, no. 3, Mar. 2025, https://doi.org/10.1007/s40808-025-02355-6.
33. Fouad, Yasser, et al. "Adaptive Visual Sentiment Prediction Model Based on Event Concepts and Object Detection Techniques in Social Media." International Journal of Advanced Computer Science and Applications, vol. 14, no. 7, Jan. 2023, https://doi.org/10.14569/ijacsa.2023.0140728.
34. Rainio, Oona, et al. "Evaluation Metrics and Statistical Tests for Machine Learning." Scientific Reports, vol. 14, no. 1, Mar. 2024, https://doi.org/10.1038/s41598-024-56706-x.
35. Choi, Yelim, et al. "Utilizing Machine Learning-based Classification Models for Tracking Air Pollution Sources: A Case Study in Korea." Aerosol and Air Quality Research, vol. 24, no. 7, Jan. 2024, p. 230222. https://doi.org/10.4209/aaqr.230222.
36. Rao, Routhu Srinivasa, Lakshmana Rao Kalabarige, M. Raviraja Holla, et al. "Multimodal Imputation-Based Multimodal Autoencoder Framework for AQI Classification and Prediction of Indian Cities." IEEE Access, vol. 12, Jan. 2024, pp. 108350–63. https://doi.org/10.1109/access.2024.3438573.
37. Barthwal, Anurag, and Amit Kumar Goel. "Advancing Air Quality Prediction Models in Urban India: A Deep Learning Approach Integrating DCNN and LSTM Architectures for AQI Time-series Classification." Modeling Earth Systems and Environment, vol. 10, no. 2, Feb. 2024, pp. 2935–55. https://doi.org/10.1007/s40808-023-01934-9.
38. Rafi, Mohammad Ariful Islam, et al. "Air Pollution Prediction and Classification With a Hybrid ANN-LSTM Model in Modern Cities: A Comparative Study." IET Conference Proceedings., vol. 2024, no. 30, Mar. 2025, pp. 580–85. https://doi.org/10.1049/icp.2025.0313.
39. Domingos, Pedro. "A Few Useful Things to Know About Machine Learning." Communications of the ACM, vol. 55, no. 10, Sept. 2012, pp. 78–87. https://doi.org/10.1145/2347736.2347755.

40. Lobo, Jorge M., et al. "AUC: A Misleading Measure of the Performance of Predictive Distribution Models." Global Ecology and Biogeography, vol. 17, no. 2, Sept. 2007, pp. 145–51. https://doi.org/10.1111/j.1466-8238.2007.00358.x.
41. Singh, Kunwar P., et al. "Identifying Pollution Sources and Predicting Urban Air Quality Using Ensemble Learning Methods." Atmospheric Environment, vol. 80, Aug. 2013, pp. 426–37. https://doi.org/10.1016/j.atmosenv.2013.08.023.
42. Zhang, Binzhe, et al. "Air Quality Index Prediction in Six Major Chinese Urban Agglomerations: A Comparative Study of Single Machine Learning Model, Ensemble Model, and Hybrid Model." Atmosphere, vol. 14, no. 10, Sept. 2023, p. 1478. https://doi.org/10.3390/atmos14101478.
43. Almaliki, Abdulrazak H., et al. "Air Quality Index (AQI) Prediction in Holy Makkah Based on Machine Learning Methods." Sustainability, vol. 15, no. 17, Sept. 2023, p. 13168. https://doi.org/10.3390/su151713168.
44. Diallo, Abdoul Aziz, et al. "Enhancing Outlier Detection in Air Quality Index Data Using a Stacked Machine Learning Model." Engineering Reports, vol. 6, no. 11, May 2024, https://doi.org/10.1002/eng2.12936.
45. Özüpak, Yıldırım, et al. "Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction." Water Air & Soil Pollution, vol. 236, no. 7, May 2025, https://doi.org/10.1007/s11270-025-08122-8.
46. Afreen, S., Bhurjee, A.K. & Aziz, R.M. Feature selection using game Shapley improved grey wolf optimizer for optimizing cancer classification. Knowl Inf Syst 67, 3631–3662 (2025). https://doi.org/10.1007/s10115-025-02340-6
47. Yaqoob, Abrar, Verma, Navneet Kumar, Rao, G. V. V. Jagannadha and Aziz, Rabia Musheer. "8 Efficient gene selection for breast cancer classification using Brownian Motion Search Algorithm and Support Vector Machine". Drug Discovery and Telemedicine: Through Artificial Intelligence, Computer Vision, and IoT, edited by Saurav Mallik, Zubair Rahaman, Soumita Seth, Anjan Bandyopadhyay, Sujata Swain and Somenath Chakraborty, De Gruyter, 2025, pp. 109-126. https://doi.org/10.1515/9783111504667-008
48. Yaqoob, Abrar, Verma, Navneet Kumar, Rao, G. V. V. Jagannadha and Aziz, Rabia Musheer. "9 A hybrid feature gene selection approach by integrating variance filter, extremely randomized tree, and Cuckoo Search algorithm for cancer classification". Drug Discovery and Telemedicine: Through Artificial Intelligence, Computer Vision, and IoT, edited by Saurav Mallik, Zubair Rahaman, Soumita Seth, Anjan Bandyopadhyay, Sujata Swain and Somenath Chakraborty, De Gruyter, 2025, pp. 127-150. https://doi.org/10.1515/9783111504667-009